UNIVERSIDAD AUTÓNOMA DE MADRID
FACULTAD DE CIENCIAS
DEPARTAMENTO DE BIOLOGÍA MOLECULAR

MULTIPARTITE VIRUSES: ORGANIZATION, EMERGENCE AND EVOLUTION

TESIS DOCTORAL

Adriana Lucía Sanz García

Madrid, 2019

MULTIPARTITE VIRUSES

Organization, emergence and evolution

TESIS DOCTORAL

Memoria presentada por
Adriana Lucía Sanz García
Licenciada en Bioquímica
por la Universidad Autónoma de Madrid

Supervisada por
Dra. Susanna Manrubia Cuevas
Centro Nacional de Biotecnología (CSIC)

Memoria presentada para optar al grado de

Doctor en Biociencias Moleculares

Facultad de Ciencias

Departamento de Biología Molecular

Universidad Autónoma de Madrid

Madrid, 2019

Tesis doctoral
Multipartite viruses: Organization, emergence and evolution, 2019, Madrid, Espana.
Memoria presentada por Adriana Lucía-Sanz, licenciada en Bioquḿica y con un máster en Biofísica en la Universidad Autónoma de Madrid
para optar al grado de doctor en Biociencias Moleculares del departamento de Biología Molecular en la facultad de Ciencias de la Universidad Autónoma de Madrid

Supervisora de tesis: Dr. Susanna Manrubia Cuevas.
Investigadora Científica en el Centro Nacional de Biotecnología (CSIC), C/ Darwin 3, 28049 Madrid, Espana.

to the reader

CONTENTS

ACKNOWLEDGMENTS

This thesis was financed by the Severo Ochoa Centers of Excellence Program (SVP-2014-068581), granted on November 16^{th} of 2014 by the Spanish Ministerio de Economía, Industria y Competitividad (Secretaría de Estado Investigación Desarrollo e Innovación). This funds also supported an internship of 3 months in the group of Prof. Eugene V. Koonin at the National Center for Biotechnology Investigation of the National Institutes of Health in Bethesda, Maryland (US).

Expenses on travelling to congresses and publishing on scientific journals were supported by the Spanish Ministerio de Economía, Industria y Competitividad and FEDER funds of the EU through grants ViralESS (FIS2014-57686-P and FIS2017-84256-P).

This thesis has been carried out with the core facilities of the National Centre of Biotechnology in Madrid (Spain), and the National Center for Biotechnology Investigation of the National Institutes of Health in Bethesda, Maryland (US).

This thesis could not have been accomplished without the contribution of the following people: Dr. Susanna Manrubia, Dr. Jacobo Aguirre, Dr. Jaime Iranzo, and the groups of Dr. Eugene V. Koonin and Dr. Mart Krupovic.

I also want to thank the scientific discussions with Drs: José Cuesta, Pablo Catalán, Juan Antonio (Toño) García and Fernando Puente-Sánchez and especially with César López-Pastrana.

RESUMEN

Los virus se encuentran entre las entidades replicativas más simples de la Tierra. Son parásitos intracelulares obligados que típicamente forman grandes poblaciones de rápida evolución. Más allá de sus altas tasas de mutación, los virus despliegan una serie de estrategias de adaptación nunca vistas en los organismos celulares. Dentro de la célula, la complementación entre genomas es una estrategia común que a menudo permite a las variantes con genomas incompletos de menor *fitness* prosperar en la población. Los virus multipartitos son el caso extremo de la complementación entre los genomas virales, ya que el genoma de estos virus está fragmentado y cada fragmento genómico es encapsidado en partículas virales independientes. Dado que todos los fragmentos genómicos tienen que coincidir en el huésped para complementarse, los virus multipartitos están obligados a luchar contra la pérdida de información genética. La necesidad de coinfección o complementación exige una alta densidad viral que es aparentemente difícil de lograr en la naturaleza. Si bien la multipartición genómica tiene una desventaja obvia como estrategia viral, no se ha llegado a un consenso sobre sus ventajas reales ¿Por qué existen los virus multipartitos?

Con el fin de contribuir al entendimiento de esta estrategia viral, cuantificamos en primer lugar la prevalencia en la Virosfera de los virus multipartitos mediante el análisis de bases de datos públicas. Entre sus características sobresalientes, encontramos que una cantidad significativa de todas las especies virales anotadas son multipartitas, y la mayoría de ellas infectan plantas. Aunque generalmente se asume que las poblaciones de virus multipartitos son viables sólo cuando la trasmisión ocurre a una alta densidad viral, la evidencia indica que son comunes los cuellos de botella, especialmente en el caso de las poblaciones de virus de plantas. Por tanto, investigamos el efecto de la fragmentación del genoma como una presión evolutiva que favorece el éxito de poblaciones multipartitas (a través de la complementación) sobre poblaciones monopartitas. Consideramos teóricamente la situación en la que se generan genomas incompletos debido a errores en la replicación de un virus monopartito parental, y determinamos cómo su persistencia se ve limitada por la densidad viral. La propagación de la infección en plantas está fuertemente condicionada por la naturaleza estructurada de los tejidos vegetales, que alivian la necesidad de complementación al favorecer una transmisión local de la infección. En consecuencia, estudiamos el caso de la fragmentación genómica en el espacio. Como resultado, en competencia espacial, las formas virales multipartitas pueden desplazar a las parentales monopartitas en condiciones ambientales menos restrictivas.

Desde un punto de vista ecológico de la Virosfera, las infecciones son el resultado de un conjunto de virus y entidades subvirales que interactúan entre sí. Los satélites son entidades subvirales que dependen de un virus ayudador para su replicación y mantenimiento. A cambio, los satélites proporcionan una nueva variedad de fenotipos de infección. Las coinfecciones con un satélite son ubicuas en las infecciones de virus de plantas, y menos comunes en otros huéspedes. Los resultados ecológicos de estas asociaciones pueden compensar el coste de la coinfección, y podrían representar un primer paso plausible hacia la multipartición genómica. Presentamos un sistema dinámico que modela la competición entre dos virus, uno de ellos asistido por un satélite, y encontramos soluciones donde prevalece el tándem virus-satélite a pesar de la necesidad de coinfectar.

Los fragmentos genómicos que constituyen los genomas multipartitos son, en principio, indistinguibles de los de otros virus; sin embargo, como virus vegetales, su origen podría estar relacionado con la expansión de los virus de ARN durante el florecimiento de las células eucariotas. La evolución de los virus eucariotas se basa en gran medida en un principio modular de construcción, impulsado principalmente por una extensa transferencia horizontal de genes. Posicionamos el origen de la multipartición genómica en la filogenia

de los virus de ARN, encontrando características sobresalientes en su evolución. La multipartición genómica ha surgido repetidamente a lo largo de la evolución y la diversificación de los virus de ARN, posiblemente impulsada por un principio constructivo que se caracteriza por un mayor intercambio de genes dentro de este grupo, en comparación con el del resto de virus de ARN.

En resumen, proponemos posibles orígenes y mecanismos evolutivos para las diferentes familias de virus multipartitos e hipotetizamos que la ventaja de la multipartición genómica se basa en su plasticidad y, al mismo tiempo, en su inevitable dependencia del contexto ecológico. Sólo explorando la interacción entre evolución y ecología podemos dilucidar qué es lo posible y qué puede ser real en las estrategias de adaptación viral.

ABSTRACT

Viruses are among the simplest replicative entities on Earth. They are obligate intracellular parasites that typically form large and fast-evolving populations. Beyond their high mutation rates, viruses deploy a number of adaptive strategies unseen in cellular organisms. Inside the cell, complementation among genomes is a common strategy that often permits variants of low fitness with incomplete genomes to thrive in the population. Complementation between viral genomes seems to be taken to its ultimate consequences in the case of multipartite viruses: the genome of these viruses is fragmented and encapsidated into independent viral particles. Since all the genomic segments have to meet in the host for complementation, multipartite viruses are bound to fight the loss of genomic information. The need for coinfection or complementation demands a high viral density that is difficult to achieve in nature. While this is an obvious disadvantage of this strategy, no consensus on its actual advantages has been reached. What is more, the main open question about multipartite viruses is why they exist at all.

In order to contribute to the understanding of this viral strategy, we quantify the prevalence of multipartite viruses by analysing publicly available databases. Amongst their outstanding characteristics, we found that a non-negligible amount of all annotated viral species are multipartite, and most of them infect plants. Although it is generally assumed that multipartitism is viable only when propagation occurs at high viral density, evidence indicates that severe population bottlenecks are common, especially for plant viruses. Therefore, we investigate the effect of genome segmentation as an evolutionary pressure that favours multipartite over monopartite populations through complementation. We consider a situation where incomplete genomes are generated through errors in the replication of a monopartite wild type virus and determine how their maintenance is constrained by viral density. Propagation of infection in plants is strongly conditioned by the structured nature of plant tissues, which alleviate the requirement of complementation by favouring local clustering. As a result, we observe that in spatial competition multipartite viral forms can displace monopartite counterparts under less restrictive environmental conditions.

Taking an ecological viewpoint of the Virosphere, infections result from an ensemble of interacting viruses and subviral entities. Satellites are subviral entities that rely on a helper virus for replication and maintenance. In return, satellites open up a new range of infection phenotypes. Virus-satellite coinfections are ubiquitous in plant infections and less common in other hosts. The ecological outcomes of these associations may compensate for the cost of coinfection and we conjecture that they represent a plausible first step towards multipartitism. We present a dynamical system for the competition between two viruses, one of them assisted by a satellite, and find solutions where the tandem virus-satellite prevails despite the requirement of coinfection.

The genomic segments composing multipartite genomes are, in principle, indistinguishable from those of other viruses; however, as a plant virus, their origin might be linked with the expansion of RNA viruses during the flourishing of eukaryotic cells. The evolution of eukaryotic viruses is highly relying on a modular constructive principle, mostly driven by extensive horizontal gene transfer. We placed the origin of multipartitism in the global phylogeny of RNA viruses and found outstanding features in their evolution. Multipartitism has repeatedly emerged along the evolution and diversification of RNA viruses, possibly driven by a constructive principle that enhances gene sharing within this group as compared to the rest of viruses.

In summary, we propose plausible mechanistic and evolutionary origins of different families of multipartite viruses and hypothesize that the power of multipartitism relies on its plasticity and, at the same time, unavoidably in an ecological context. It is only by

exploring the interplay between evolution and ecology that we can elucidate the possible and the actual in viral adaptive strategies.

INTRODUCTION

I.1 What is a virus?

The first question to address in a thesis related to viruses has to be *"What is a virus?"* (Roossinck, 2016; Knipe and Howley, 2007). A quick look to the The Oxford English Dictionary defines a virus as:

> "an infective agent that typically consists of a nucleic acid molecule in a protein coat, is too small to be seen by light microscopy, and is able to multiply only within the living cells of a host."

The definition is a slightly vague taking into account that the characteristics proposed for a virus are shared by other unrelated organisms. First, it is said that viruses are infectious agents, as many bacteria, fungi and protozoa. The term "infectious agent" is commonly associated to *disease-causing organism*. Viruses are related with disease, but the reality is that many viruses do not cause any harm.

The second feature is just half true. A prototypic virus has an icosahedral shell protecting the viral genome, but it uses to come in various shapes – filaments, rods, bottle-like, spider-like. Some viruses also have an external lipid membrane known as envelope, which surrounds the entire capsid; others, like narnaviruses or endornaviruses, simply don't have a capsid at all.

Third, it is said that they are too small to be seen by light microscopy – whose resolution is about a half of the wavelength of visible light $\sim 0.2\mu m$. The average size of most viruses is around 0.01 μm well beyond the resolution limit of a light microscope. Therefore, viruses are small, much smaller than a bacteria or a blood cell, around 100 and 1000-fold smaller Figure I.1. To make a graphical example we can compare a cell infected by a virus with a human being attacked by a cockroach. Nonetheless, there are giant viruses, with

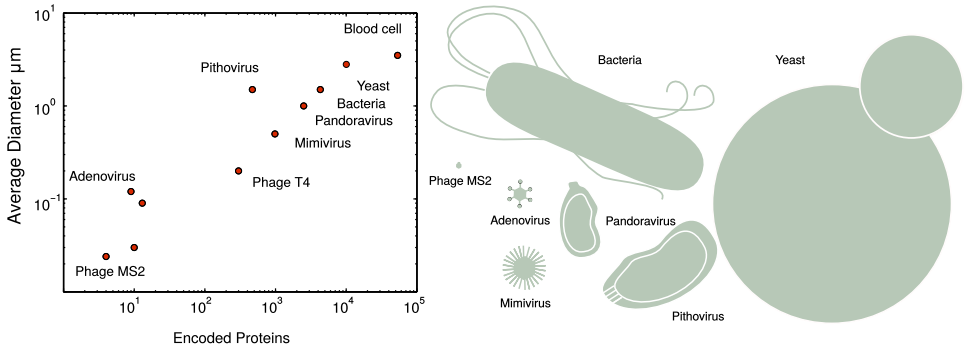

Figure I.1: **Allometric relationship between physical size and genome size for several viruses and cellular organisms.**
The graphic shows a comparison of diameter size and the number of encoded proteins for several viruses and cellular organisms. The companion picture shows relative sizes of some of the organisms included in the graphic.

sizes reaching almost $1.5\mu m$ that can actually be seen by light microscopy. Mimivirus and pandoravirus are giant viruses that redefine the concept of how small a virus can be (Scola et al., 2003; Philippe et al., 2013). These viruses are only 5 times smaller than archaea, their natural host.

Last feature mentioned is perhaps the most significant because it is shared by all viruses, as no known virus is able to multiply without parasitizing a living cell. However, not only viruses are mandatory intracellular parasites. There are fungi, protozoa and bacteria that cannot replicate without a host such as, chlamydia, rickettsia, or leishmania (Knipe and Howley, 2007).

Yet, a deeper understanding on viruses has to highlight the differences of viruses from cellular life forms. The non-cross line for viruses is that of having their own translation machinery, therefore they depend on the cellular one to express their genetic information. There are viruses that encode only one protein —narnavirus— but others like the giant pandoravirus encode a complete metabolism with thousands of proteins involved **Figure I.1**. The genetic information of viruses is coded in six different types of nucleic acids – single or double stranded RNA or DNA, of positive or negative polarity– according to the Baltimore classes (Baltimore, 1971), in contrast with the sole genetic molecule for cellular life forms, the dsDNA. Another remarkable difference of viruses is the diversity in their lifestyles as opposed with the uniformity of the cellular cycle.

Viruses are everywhere. Every living organism –from bacteria to humans– when studied, are infected by viruses. We cannot image a cellular form free of viral infection. Viruses are such an ubiquitous parasite, that we cannot think on life as we know it without their action. We find viral genes in genomes of living organisms and vice versa, in a way that we cannot discern the evolution of live without the action of viruses, as they are major vehicles of horizontal gene transfer in the evolution of life. Probably, viruses existed since living cells first evolved, ultimately meaning we cannot understand cellular life, without the impact of viral infections, their role in the ecosystems as real players of evolution.

Consequently, in this thesis we define a virus as:

> "a non-cellular microscopic parasite, that lacks the capacity to translate their genetic information into proteins, infects all types of life forms and consists of a genetic material based on DNA or RNA usually surrounded by a protein and/or membrane coat."

The mechanisms by which viruses evolved together with cellular life (Koonin and Dolja, 2013), how they adapt in a continuous arms race towards parasitism (Koonin, 2016), the strategies they display to find a way to colonize novel niches to infect, are as of yet major unknowns from the viewpoint of evolution (Koonin et al., 2006). However, many aspects of the variety of genetic strategies, viral lifestyles, genome complexity and phylogeny and global ecology of viruses have been extensively studied and set a basis to infer the big questions of virus evolution. While a major challenge for evolutionists is to address these questions, an ambitious goal of this thesis is to explore a particular virus strategy which is one of the most puzzling ones found in the Virosphere.

I.2 What is a multipartite virus?

This thesis focuses on a particular class of viruses known as multipartite viruses first described in the 1960 decade (Lister, 1966; van Kammen, 1967). Several names have been ascribed to multipartite viruses in the literature, such as coviruses, multicomponent viruses, multiparticle or multicompartment viruses. The particularity of these viruses is intrinsically related to their genome configuration and their transmission style. We are considering three main types of viral genome configurations: non-segmented or monopartite, segmented and multipartite viruses. Non-segmented and segmented viruses transport all the genomic information needed to complete the viral cycle inside a unique viral particle whereas multipartite viruses distribute their chromosomes into two or more virus particles as shown in **Figure I.2**. Coinfection is therefore a requirement for survivability of multipartite viruses, a requirement that is absent in the rest of genome configurations. This is a viral strategy bounded to fight the loss of genomic information, as viral particles containing the different genomic segments are independently transmitted. There is a need for co-infection or complementation that demands a high viral density —multiplicity of infection (MOI)— not attainable, or at least very difficult to achieve in most of the cases (Reanney, 1982; Gutiérrez et al., 2010).

There are many more open questions than certainties in our understanding of multipartite viruses (Nee, 2000; Sicard et al., 2016; Holmes, 2016). Among all, the main puzzle is why multipartite viruses do exist at all (Wu et al., 2017). Despite our current lack of knowledge on the mechanisms that may endow multipartite viruses with an adaptive advantage that compensates for their apparently weird and costly lifestyle, there is no alternative but beginning by assuming that multipartitism is a stable evolutionary strategy, *sensu* Maynard-Smith (Maynard Smith, 1972). The following section will introduce the basics of a multipartite virus lifestyle, its particularities and commonalities with non-segmented and segmented viruses in order to introduce the terminology that will be discussed in part III of Results and Discussion.

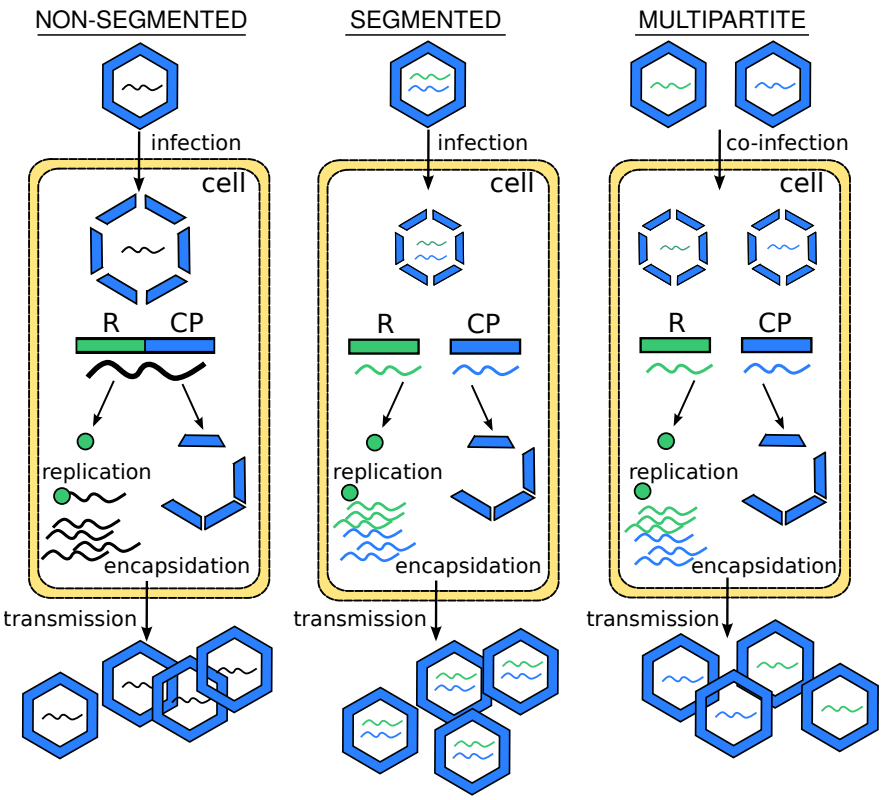

Figure I.2: **Infection cycles attending to genome configuration.**
Non-segmented (or monopartite) viruses transmit all the genomic information in a single genetic molecule contained in a single capsid coat. When they infect a cell, the main functions of replication, R, and encapsidation, CP, are codified in the genetic molecule that has just entered the cell. Segmented viruses transmit all the genomic information in two to several genetic molecules contained in a single capsid coat. When they infect a cell, all the necessary functions to complete the infection cycle are codified in the various genetic molecules that have just entered the cell. Multipartite viruses transmit the genomic information in two to several genetic molecules contained in independent capsid coats. Two to several viral particles must co-infect the same cell to ensure that all genomic functions are present to complete the infection cycle. Complementation is a constraint only reserved to a multipartite genome configuration.

I.3 The multipartite lifecycle

Multipartite viruses have shown to be really successful plant viruses. Despite this preference might be a bias in the host-range due to biased sampling, we will assume that plants offer adequate conditions for the maintenance of multiparticism.

Possibly as a consequence, the nucleic acids linked to multipartitism are the same than those of other non-segmented and segmented plant viruses, with only one exception: plant retroviruses, which are all non-segmented. Multipartite genomes are molecules of double and single stranded RNA of positive and negative polarity (dsRNA, +RNA, -RNA),

and in single stranded DNA (ssDNA). The nucleic acid is an important determinant of the infection cycle as it imposes restrictions in the sequence of steps to complete the infection (Ahlquist, 2006) and the proteins needed to interact with the different cellular components of the host. Whereas the nucleic acid strongly affects the infection cycle, there is no evidence that the genome configuration plays any role on it, and in principle viral infection cycles of multipartite, segmented and non-segmented viruses of a particular class of nucleic acid are undistinguishable.

While RNA viruses limit their activity to the cytoplasm, with several exceptions (Cros and Palese, 2003; Krichevsky et al.), DNA viruses have to enter the nucleus for replication and transcription, a step that requires the assistance of viral proteins that allow the export and import through the nucleus (Whittaker and Helenius, 1998).The +RNA viruses have the simplest infection strategy for replication of their genome and expression of their genes, since the same molecule serves as genetic material and translation template. Right after the genome is released from the capsid, it is translated into viral proteins as an initial step in the infection cycle. In contrast, dsRNA and -RNA viruses need a replicase to transcribe the genome into viral messenger RNAs (mRNA), as first step before translation. Therefore, the replicase protein has to be transmitted together with the genome inside the virion.

Although viruses with +RNA genomes are able to start the infection cycle without the assistance of any protein, some of them produce sub-genomic messenger RNAs (sgRNA) later in the infection cycle, that results in a second round of translation. The sgRNAs use to be shorter mRNAs that derive from a replication intermediate molecule of dsRNA, the precursor of genomic +RNA. This intermediate molecule of dsRNA has to be protected from the cytoplasm RNA silencing surveillance and preserved from degradation inside vesicles or capsids. Viral complexes or factories made up from cellular membranes are observed in infections of +RNA viruses (Laliberté and Sanfaçon, 2010; Laliberté and Zheng, 2014). Similarly, dsRNA viruses never release their genome from the capsid in order to avoid the contact with the cytoplasm (Ahlquist, 2006). Instead a dsRNA intermediate, -RNA viruses use a different mechanism for replication and transcription since -RNA molecule is always protected by a nucleocapsid, forming a ribonucleoprotein complex (RNP) (Reguera et al., 2016). The polarity of the -RNA genomes imposes transcription as the obligatory first step in the virus gene expression programme, similar to dsRNA viruses. Thus, viral mRNAs are first produced from the parental RNPs while RNA replication intermediates are generated at a later stage when viral nucleocapsids have been produced (Ortín and Martín-Benito, 2015).

Multipartite viruses can also have ssDNA genomes. While the mechanism of translocation of the genome to the nucleus is unknown, they do not package any additional proteins inside the capsid. However, they must encode viral proteins to mediate the transport through the nucleus of additional viral factors to initiate the replication through the rolling circle mechanism (Krichevsky et al.).

A detailed explanation of the viral cycles according to the nucleic acids composing the genomes aforementioned are in the schemes of **Figure I.3**.

Another factor that modulates the infection cycle is the host. Plants are particular hosts because the majority of their tissues are connected by specific channels named plasmodesmata through which cells share nutrients, transmit information, and in a context of infection, serve as treadmills to spread viral infections (Niehl and Heinlein, 2010; Dall'Ara et al., 2016). Most plant viruses are able to control these channels by means of movement proteins that allow the movement of viral particles, proteins or genomes from one cell to its neighbours (Kaido et al., 2011; Kawakami et al., 2004). A scheme of the molecular mechanisms that plant virus display to spread the infection from cell tissues to the

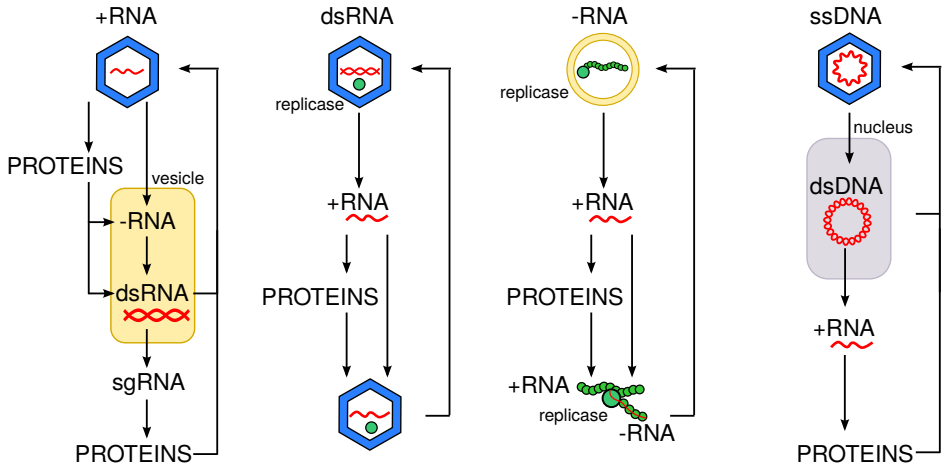

Figure I.3: **Infection cycles of multipartite viruses constrained by genetic molecule.** Once the virus particle gets disassembled in the cytoplasm, the genome of +RNA is directly translated into viral proteins by cellular ribosomes. Then, the virus controls cellular membranes and produce replication complexes inside vesicles where anti genomic -RNA strands are produced directly from the genomic +RNA and a dsRNA intermediate of replication is isolated from the cytoplasm. A second round of translation takes place from sgRNAs and structural proteins are produced massively to generate capsids where the genomic +RNA molecules are encapsidated. Viruses with dsRNA do not get disassembled after the entrance in the cytoplasm and the replicase inside the capsid transcribe dsRNA into viral messenger +RNA and genomic +RNA that exit the capsid and go to the cytoplasm. After translation, new capsids are assembled and genomic +RNA is encapsidated together with the replicase. Then, genomic dsRNA is produced inside the new capsids. Similar to dsRNA viruses, viral messenger +RNA and anti genomic +RNA are produced from -RNAs by viral replicases that are co-packaged with the genome. The genetic molecule of -RNA viruses and anti genomic +RNAs are always wrapped by a nucleocapsid forming a riboproteic filament. During replication nucleocapsids are displaced by the replicase from the template genetic strand of +RNA and new genomic molecules of -RNA are wrapped by newly synthesized nucleocapsids. Then the genomic ribonucleocapsids are encapsidated together with the replicase. After uncoating, the viral ssDNA genome penetrates into the nucleus and is converted into dsDNA with the participation of cellular factors. dsDNA transcription produces viral mRNAs and translation of viral proteins. Replication occurs by rolling circle mechanism producing ssDNA genomes. These newly synthesized ssDNA can either be converted into dsDNA and serve as a template for transcription/replication or be encapsidated and released though cell lysis.

rest of the plant are depicted in **Figure I.4**. There is a strong evidence that during the movement, multiple copies of a genome variant, probably inside genome complexes or viral factories (Kawakami et al., 2004; Cotton et al., 2009), are collectively delivered to the neighbour cells, a process that assures that all the genetic segments of a multipartite virus pass collectively to the next cell without loss of genetic information (Miyashita and Kishino, 2010). In addition, several studies shown that viral proteins and sgRNAs actually

diffuse long distances through the plasmodesmata from the initial infected cell (Heinlein et al., 1998; Sicard et al., 2019). This process could, in principle, complement any function *in trans* to progress the course of infection in a cell where that specific genomic segment was not present (Sicard et al., 2019) **Figure I.4.C**.

To efficiently travel to other leaves and cause systemic infections, plant viruses have to cross the vascular tissues, where they still move from one cell to another through wider channels that connect vascular cells. There are plant viruses that cannot control the plasmodesmata and are they restricted to vascular tissues and phloem of the plant with no access to the leaves (Rasheed et al., 2006; Stewart et al., 2012).

Once a virus establishes an infection in a plant cell it usually impairs the entry of other related viruses to the same cell, a process called superinfection exclusion. This mechanism is based on the capacity of the viruses to control the plasmodesmata and other plant cell channels, by blocking them when the infection is established in the cell. It is thought that cell-to-cell movement is strongly affected by purifying selection (Zwart et al., 2014), and out of the $6 \cdot 10^7$ genomes that are approximately generated during a replication cycle in a cell (Nixon), the actual number of genetic variants that are available for transport are between one and two (Miyashita and Kishino, 2010; Tromas et al., 2014), and depends mainly on the timing of movement in relation to the viral replication cycle (Tilsner and Oparka, 2012). Cell-to-cell movement and superinfection exclusion are principal determinants for the low genetic variability observed in plant viruses (Dunham et al., 2014), and result in a phenomenon called spatial clustering: a pattern that appears in infected leaves and veins when different virus biomarked strains are spatially segregated (Gutierrez et al., 2012). Whereas superinfection exclusion imposes a severe bottleneck during virus propagation in the plant (Zwart and Elena, 2015), it seems to be permissive to the need for complementation of the genetic segments of multipartite viruses. Collective cell-to-cell movement and viral protein sharing are possibly a way to overcome the constraint of complementation for multipartite viruses, and a key factor for the success of multipartitism as a strategy of plant infection compared to other hosts. All together, a structured propagation mode seems to favour multipartite viruses (Aguirre and Manrubia, 2008), and plants offer the perfect environment for this to take place.

An exceptional feature of multipartite viruses is the observation that particles containing the different genetic segments are not equally abundant after plant infection. The differences of accumulation levels of the genome segments have been reported in laboratory conditions for several unrelated multipartite families (Grigoras et al., 2009; Sánchez-Navarro et al., 2013; Sicard et al., 2013; Wu et al., 2017). Interestingly, regardless of the initial infection ratio, the relative frequencies amongst the genomic segments reach a constant composition that could be host-specific (Sicard et al., 2013). Genetic imbalance may occur at some point during the interaction with the host and, although the mechanism behind is still unknown, it is likely to be linked to plant transportation and to the complementation of functions *in trans* previously explained. Several experiments pointed out differences in the mechanisms for systemic transportation of each segment of a tripartite virus. The systemic movement of two of the genomic segments involve genomic RNA–RNA interactions but is independent of the third segment which interacts with other viral and host factors for transportation (Torrance et al., 2009). It may be obvious that independent mechanisms for long-distance movement in the plant will be reflected into differences in the accumulation levels of each segment, although further data is needed to confirm this insight. Another recent report indicates that different gene segments are preferentially replicated in each cell when they co-localize with the replicase protein, even though the genetic segment that encodes the replicase is absent. This way of replication might correspond with a winner-

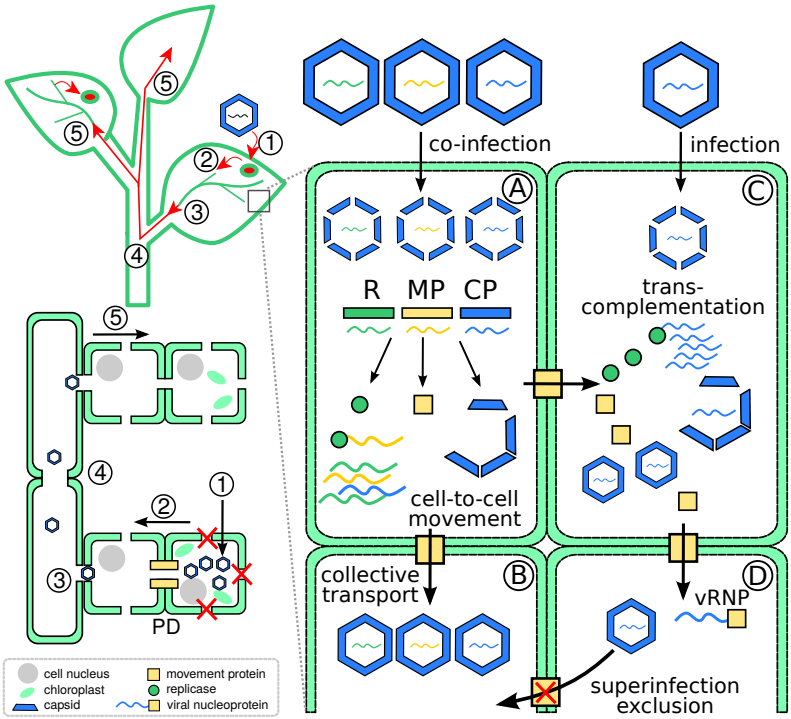

Figure I.4: **General view of virus cell-to-cell and long-distance movement in plant tissues.**
(1) Viruses enter the plant epidermis through physical inoculation by insect vectors or by lesions in the leaves forming a local spot. Once inside a cell, virions are disassembled for replication, and expression of the main functions encoded in the viral genome: replication, R, encapsidation, CP, and cell movement, MP (A). (2) Movement proteins, mediate the transport of virions (B) or viral proteins (C) from cell-to-cell. They can enlarge cellular channels called plasmodesmata, PD, and allow the circulation of viral particles or alternatively, sometimes associated to cellular factors, they can interact with the viral genome to form transport complexes (vRNP) (D). Viral proteins, transported from cell-to-cell, can complement functions *in trans*, facilitating the infection in cells where the infection has not been fully established (C). Superinfection exclusion prevents a plant cell to get infected by surrounding viruses, until the current infection cycle is not finished (B). Viruses continue replicating inside every infected cell until they reach the vein tissue for long-distance movement (3) before being finally getting released into systemic tissues (4) where they move freely without the assistance of MPs. Transported in the phloem, viruses enter upstream tissues to start new infection sites (5).

takes-all strategy that could be behind segment unbalance. Task allocation might be a way to optimize segment production in multipartite viruses, but this hypothesis has not been proved yet.

In order to move from one plant to another most plant viruses, including multipartite, are transmitted by an insect vector that feeds on the plant (Power, 2000; Whitfield et al., 2015; Dietzgen et al., 2016). Other plant pathogens have shown to serve as vehicles of plant virus transmission such as other arthropods (Nault, 1997), nematodes and fungi, but horizontal

transmission through seeds, or cell division is the second most frequent transmission form, causing persistent infections in plants (Roossinck, 2010). Vector transmission requires some degree of specificity and in general specific viral proteins mediate the interaction of the virus with its vector (Rahim et al., 2007; Deshoux et al., 2018) and vice versa, which may be the result of a long-term co-evolution of virus with the vector. In fact, there is a high specificity of most plant viruses by their vectors, since only a few related insect species are usually capable of transmitting the virus, whereas the host-range can include several species of different genera or even plant families (Power and Flecker, 2010; Lefeuvre et al., 2019). The degree of specificity of the virus by its vector results in differences on transmission efficiency. Viruses that only interact with vector mouth parts are poorly transmitted and only around 0.5 to 2 virus particles are transferred to susceptible plants (Moury et al., 2007; Betancourt et al., 2008). However, viruses which enter the foregut or even replicate in the vector are transmitted in greater amounts (Gallet et al., 2018b). The overall estimated number of genomes transmitted by one insect vector is very narrow and, in principle imposes a severe bottleneck for the virus population. Although the benefits of vector transmission are numerous, because these insects usually travel long distances and feed on different plant species providing many colonization options, low transmission numbers impose a major constraint for plant virus evolution (Power, 2000; Zwart and Elena, 2015; Lefeuvre et al., 2019). Many vector insects propagate in plagues, a strategy that may result in a larger number of virus particles entering the host, but all together, it is remarkable that these severe bottlenecks are compatible with the persistence of multipartite viruses. Collective infection strategies may also contribute to the transmissibility of multipartite viruses, including the possibility that several viral particles aggregate forming a collective infection unit, or a versatile ploidy of the virions that permits the occasional encapsidation of multiple copies of the genome in one capsid. However, these transmission strategies have not been confirmed in multipartite viruses (Richards and Tamada, 1992), but in several segmented (Lago et al., 2016; Galasso, 1967; Rager et al., 2002) and non-segmented viruses (Bald and Briggs, 1937; Beniac et al., 2012).

Plant virus populations have to cope with severe bottlenecks, both in the spread of the virus from plant-to-plant and from cell-to-cell within the host. Considering that such reduced populations are subjected to a strong genetic drift, the adaptation not only to their hosts, but also to the vector required for their transmission is particularly constrained. Although certain plant particularities could favour the multiplication of multipartite viruses, it is still hard to imagine how multipartite virus populations adapt to the various imposed evolutionary pressures and how they maintain genome integrity.

I.4 Overview of this thesis

This thesis is motivated by the intriguing evolution of multipartite viruses. We present an integrated overview of multipartite viruses and their apparent costly lifestyles. We analyse in depth the hypotheses about their emergence and persistence, as well as the evolutionary mechanisms that give rise to the transition to multipartitism. We explore the origin of multipartitism in the context of RNA virus evolution and the construction mechanisms of this genetic configuration.

The Chapter 1 presents an exhaustive search and integration of data from different available sources that serves to elaborate a detailed reference guide of multipartite and segmented viruses. The guide contains data on prevalence, families and genera, host range, number of genomic segments of all multipartite and segmented species found to date. The

outstanding features of multipartite compared to segmented viruses are also discussed in this chapter.

The Chapter 2 presents an extensive review of the empirical observations of multipartite viruses, the proposed advantages of multipartitism to overcome the cost of multiple infection, together with theoretical investigations of multipartite infection. Additionally, this chapter describes a theoretical model for the evolution of multipartitism motivated by the most relevant empirical observations.

The Chapter 3 discusses the ecological consequences of virus-satellite associations that may compensate for the cost of coinfection of both entities. That may represent a key role towards the emergence of a multipartite species. An original model is introduced and discussed in order to understand the ecological effects of those associations in a framework of viral competition. The implications on the evolution of multipartitism are also analysed.

The Chapter 4 discusses the mechanisms governing the transition to multipartitism and presents an hypothetical scenario for the transition towards multipartitism. The origin of multipartitism is contextualized for the different Baltimore classes were multipartitism has been found. We focus in depth on the origin of multipartitism and genome segmentation in the evolution of RNA viruses. We present key evolutionary features of multipartite and segmented viruses in the context of the global RNA phylogeny. This chapter is part of the work with the groups of Eugene V. Koonin and Mart Krupovic, and is contributed by the collaboration of Yuri I Wolf, Darius Kazlauskas, Jaime Iranzo, Jens H Kuhn, Mart Krupovic, Valerian V Dolja, Eugene V Koonin.

The Chapter 5 is the last chapter dedicated to a general discussion of the thesis, and it emphasizes on the perspectives of this work.

PART I

OBJECTIVES

The global objective of this thesis is to advance in the understanding of genome multipartition in virus evolution. The general objective of the thesis is divided into several specific objectives.

1. Build a reference guide of multipartite viruses, to put them into context in relation to the rest of the Virosphere in terms of outstanding features, prevalence and representative families.

2. Understand the principal mechanisms involved in the emergence of multipartite viruses.

3. Study of the origin of multipartite viruses in the context of the evolution of RNA viruses.

PART II

METHODOLOGY

0.5 Database management for constructing the multipartite and segmented datasets

We initially look for multipartite and segmented families candidates retrieving information from ViralZone, a database available at the website http://viralzone.expasy.org. The ViralZone project is handled by the virus program of the SwissProt group in the Swiss Institute of Bioinformatics. ViralZone is an SIB Swiss Institute of Bioinformatics web-resource for all viral genus and families, providing general molecular and epidemiological information, along with virion and genome images. Each genus or family page gives an overview of the basic biological information, virion and genome structure and composition, brief lifecycle and an easy access to UniProtKB/Swiss-Prot viral protein entries. Since data is only available online, information was retrieved pseudo-manually. In particular, we were interested in the taxonomy browsed by the genome statistics where we retrieved family candidates with segmented genomes. We set a preliminary dataset of segmented and multipartite families and genera. We retrieved individual information about virion structure, number of genome segments and mode of transmission for each family. This dataset was subsequently curated by looking at the literature to corroborate the preliminary candidates

and by manually adding new families and unassigned genera to obtain a final dataset of multipartite and segmented families and genera.

The next step was to calculate the number of genera and species within each viral family. We used the latest version of the International Committee on Taxonomy of Viruses database (ICTV) that was downloaded from the website https://talk.ictvonline.org/ in a table format. The ICTV provides a universal virus taxonomy and a classification scheme for most of the viral species described to date that is supported by verifiable data from nucleotide sequences and expert consensus. We used the ICTV database to quantify the number of species and genera within each family and genera using simple python or C routines that allow file input/output and string comparison.

We assigned a host for each of the species in the dataset using a virus host database available at the website https://www.genome.jp/virushostdb/. The Virus-Host DB is an original database produced by the Laboratory of Chemical Life Science. The data can be downloaded from an FTP server in table format, where data is represented in the form of pairs taxonomy identifiers for viruses and their hosts. The taxonomy identifiers are those of the of National Center for Biotechnology Information. We obtained the taxonomy identifiers for each of the virus species in the dataset, retrieving the information directly from the National Center for Biotechnology Information. We used the e–utilities an query and database system at the National Center for Biotechnology Information (NCBI). The e–utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data.

0.6 Analytical solutions and stability analysis of the model of viral competition

We obtain the stationary solutions –fixed points– (H^*, X^*, Y^*) of the system by equating to zero the ordinary differential equations describing the model ($\dot{H} = 0$, $\dot{X} = 0$ and $\dot{Y} = 0$ and obtaining the solutions H^*, X^* and Y^*. Conditions for existence and non-negativity of all variables have to be fulfilled for those solutions to be meaningful (being either a stable or an unstable fixed point). Four stable non-negative equilibria solutions are found:

$$(H^*, X^*, Y^*) = \left(\frac{g}{d}, 0, 0\right) \tag{0.1}$$

$$(H^*, X^*, Y^*) = \left(\frac{d+d_x}{p_x}, \frac{g}{d+d_x} - \frac{d}{p_x}, 0\right) \tag{0.2}$$

$$(H^*, X^*, Y^*) = \left(\frac{d+d_y}{p_y}, 0, \frac{g}{d+d_y} - \frac{d}{p_y}\right) \tag{0.3}$$

$$(H^*, X^*, Y^*) = \left(\frac{d+d_x}{p_x}, X^*, \frac{g}{d+d_y} - \frac{d}{p_y} - \frac{p_x}{p_y}X^*\right) \tag{0.4}$$

In order to simplify the mathematical formulation we are going to define the reproductive ratio, or reproductive success, R_i which measures the ability of either virus to invade the host population.

$$R_i = \frac{p_i}{d+d_i} \quad with \quad i \in \{x, y\} \tag{0.5}$$

The ratio d/g can be seen as a measure of the replacement or turnover time of healthy hosts.

In the solution of eq. (0.1) none of the competing virus survives and the host population $H^* = g/d$ is constant at equilibrium. The solutions in eqs. (0.2) and (0.3) correspond to the survival of only one of the viruses infecting the population of hosts, thus entailing the extinction of the other one. As a result of the symmetry of the system, eqs (3.2) and (3.3) are analogous. Finally, the degenerate solution eq. (0.4) does not determine unequivocally the variable X^*, indicating coexistence of the two competing viruses.

0.6.1 Stability analysis of the model of viral competition

The conditions for stability of the solutions are given by the sign of the eigenvalues of the Jacobian matrix associated to the system. When all eigenvalues are negative, the fixed point is stable. When one of the eigenvalues is zero and the rest are negative, the fixed points degenerate into a continuous set of quasi-stable solutions. The Jacobian matrix, J, is the matrix of all first-order partial derivatives of a function or set of functions, and in our system takes the form:

$$J = \begin{pmatrix} \frac{\partial K_1}{\partial H} & \frac{\partial K_1}{\partial X} & \frac{\partial K_1}{\partial Y} \\ \frac{\partial K_2}{\partial H} & \frac{\partial K_2}{\partial X} & \frac{\partial K_2}{\partial Y} \\ \frac{\partial K_3}{\partial H} & \frac{\partial K_3}{\partial X} & \frac{\partial K_3}{\partial Y} \end{pmatrix} = \begin{pmatrix} -d - p_x X^* - p_y Y^* & -p_x H^* & -p_y H^* \\ p_x X^* & p_x H^* - (d + d_x) & 0 \\ p_y Y^* & p_y H^* - (d + d_y) & 0 \end{pmatrix} \tag{0.6}$$

Where the functions $K_{1,2,3}$ are:

$$\begin{aligned} K_1 &= g - dH - p_x XH - p_y YH \\ K_2 &= p_x XH - (d + d_x)X \\ K_3 &= p_y YH - (d + d_y)Y \end{aligned}$$

We evaluate the J matrix of the system eq.(0.6) for every fixed point in eqs.(0.4-7) as $J(H^*, X^*, Y^*)$ and find, analytically when possible, the roots or eigenvalues, λ_i, of the characteristic polynomial $p(\lambda) = det(J(H^*, X^*, Y^*) - \lambda \mathbb{I})$ by making $p(\lambda) = 0$. Solutions are stable if the λ_i are either negative or equal to zero. The conditions of existence and stability are given for each solution and are discussed in depth in Chapter 3:

1. Conditions for stability, existence and non-negativity of the solution in eq. (0.1) $(H^*, X^*, Y^*) = \left(\frac{g}{d}, 0, 0\right)$ are:

   $$R_x < d/g \text{ and } R_y < d/g$$

   The conditions above indicate that if the growth of healthy hosts is sufficiently low, it prevents viral invasion of the population. Without the possibility to replicate in new susceptible hosts, both viruses get extinct.

2. Conditions for stability, existence and non-negativity of the solution in eq. (0.2) $(H^*, X^*, Y^*) = \left(\frac{d+d_x}{p_x}, \frac{g}{d+d_x} - \frac{d}{p_x}, 0\right)$ are:

   $$R_x > d/g \text{ and } R_x > R_y$$

A virus with a replicative success above that of its competitor and also higher than the turnover time of a healthy host will invade the population.

3. Conditions for stability, existence and non-negativity of the solution in eq. (0.3) $(H^*, X^*, Y^*) = \left(\frac{d+d_y}{p_y}, 0, \frac{g}{d+d_y} - \frac{d}{p_y}\right)$ are:

$$R_y > d/g \text{ and } R_y > R_x$$

4. Conditions for stability, existence and non-negativity of the solution in eq. (0.4) $(H^*, X^*, Y^*) = \left(\frac{d+d_x}{p_x}, X^*, \frac{g}{d+d_y} - \frac{d}{p_y} - \frac{p_x}{p_y}X^*\right)$ are:

$$R_x = R_y > \frac{d}{g} \text{ and } \frac{g}{d+d_x} - \frac{d}{p_x} > X^* > 0$$

The conditions for coexistence of both viral types are very stringent in parameter space, since only when the replicative successes of both viruses are equal and larger than the turnover time of healthy hosts can coexistence be an stable outcome of the system.

0.7 Results of the model of viral competition assisted by a satellite

We obtain the fixed points (H^*, X^*, Y^*, S^*) of the system by making the ordinary differential equations equal to zero ($\dot{H} = 0$, $\dot{X} = 0$, $\dot{Y} = 0$ and $\dot{S} = 0$) and obtaining the values of H^*, X^*, Y^* and S^* that verify them. The same way we did in the previous model, we only consider solutions that exist and are non-negative for all variables.

For $S^* = 0$ four stable solutions are found analogous to solutions in eqs. (0.1–0.4) of the previous model:

$$(H^*, X^*, Y^*, S^*) = \left(\frac{g}{d}, 0, 0, 0\right) \tag{0.7}$$

$$(H^*, X^*, Y^*, S^*) = \left(\frac{d+d_x}{p_x}, \frac{g}{d+d_x} - \frac{d}{p_x}, 0, 0\right) \tag{0.8}$$

$$(H^*, X^*, Y^*, S^*) = \left(\frac{d+d_y}{p_y}, 0, \frac{g}{d+d_y} - \frac{d}{p_y}, 0\right) \tag{0.9}$$

$$(H^*, X^*, Y^*, S^*) = \left(\frac{d+d_x}{p_x}, X^*, \frac{g}{d+d_y} - \frac{d}{p_y} - \frac{p_x}{p_y}X^*, 0\right) \tag{0.10}$$

For $S^* \neq 0$ three additional solutions are found:

- Extinction of X population:

$$\begin{aligned} H^* &= \frac{2gp_s}{K \pm \sqrt{K^2 + 4gp_s\left(-dp_s - p_y p_{sy} + \frac{(d+d_y)(p_{sy}+p_Y)}{d+d_s}\right)}} \\ X^* &= 0, \\ Y^* &= \frac{d+d_s}{p_s} - \frac{p_{sy}}{p_s}H^*, \\ S^* &= -\frac{(d+d_y-p_yH^*)(d+d_s-p_{sy}H^*)}{p_s(d+d_s-(p_sy+p_Y)H^*)}. \end{aligned} \tag{0.11}$$

where $K = p_s(d+g) + p_y(d+d_s) - (d+d_y)(p_{sy}+p_Y)$.

Co-existence of Y and S populations in eq.(0.11) indicates that when S*¿0 free-satellite populations Y*¿0 can still infect without the assistance of the satellite, and in addition, when Y*=0 automatically implies that S*=0 by eq. (3.8).

- Co-existence of all types $(H^*, X^*, Y^*, S^*) \neq 0$ has two different solutions, a stable fixed point eq.(0.12) and a degenerate solution of type $(H^*, X(S^*), Y^*, S^*)$ for any value of $S^* > 0$ eq.(0.13):

$$\begin{aligned}
H^* &= \frac{d+d_x}{p_x}, \\
X^* &= \frac{g}{d+d_x} - \frac{d}{p_x} - \frac{p_y p_Y (d+d_x)}{p_x p_s} + \frac{(p_{sy}+p_Y)(d+d_y-p_y H^*)(d+d_s-p_{sy}H^*)}{p_x p_s (d+d_s-(p_{sy}+p_Y)H^*)}, \\
Y^* &= \frac{d+d_s}{p_s} - \frac{p_{sy}}{p_s} \quad H^*, \\
S^* &= -\frac{(d+d_y-p_y H^*)(d+d_s-p_{sy} \quad H^*)}{p_s(d+d_s-(p_{sy}+p_Y) \quad H^*)}.
\end{aligned} \tag{0.12}$$

$$\begin{aligned}
H^* &= \frac{d+d_x}{p_x}, \\
X^* &= \frac{g}{d+d_x} - \frac{d}{p_x} - \frac{p_Y(d+d_y)}{p_x p_s} + \frac{p_{sy}+p_Y}{p_x} \quad S^*, \\
Y^* &= \frac{p_Y}{p_s}\frac{d+d_x}{p_x}, \\
S^* &> 0.
\end{aligned} \tag{0.13}$$

We are going to extend our definition of reproductive success in eq. (3.4) to include the reproductive ratio of the tandem satellite-virus in order to simplify the mathematical formulation. We had defined the productive success R_i as the ability of either virus (and virus-satellite) to invade the host population.

$$R_i = \frac{p_i}{d+d_i} \quad with \quad i \in \{x, y, sy, c\} \quad where \quad p_c = p_{sy} + p_Y \tag{0.14}$$

The meaning of R_{sy} differs from that of R_c. Let us assume that R_{sy} is the reproductive success of the association between the virus and the satellite when they are jointly co-transmitted (coinfection), while R_c is the replicative success of the tandem virus-satellite, despite a potential loss of the satellite during transmission. This model assures R_c is always larger than R_{sy} because intuitively the possibility of an independent transmission is larger than a coinfection of the two entities.

0.7.1 Stability analysis of the model of viral competition assisted by a satellite

The conditions of existence and non-negativity are those ensuring that every solution (H^*, X^*, Y^*, S^*) are positive or zero. The conditions of stability (translating into certain relationships to be fulfilled by the system parameters) guarantee that all eigenvalues

of the Jacobian matrix (described in the previous section) eq.(0.6) evaluated at a certain fixed point of the system, are negative. When one of the eigenvalues is zero and the rest are negative, the fixed points degenerate into a continuous set of quasi-stable solutions.

The Jacobian matrix for the model of viral competition assisted by a satellite is:

$$\begin{pmatrix} -d - p_x X^* - p_y Y^* - p_c S^* & -p_x H^* & -p_y H^* & -p_c H^* \\ p_x X^* & p_x H^* & 0 & 0 \\ p_y Y^* + p_Y S^* & 0 & p_y H^* - D_y - p_s S^* & p_Y H^* - p_s Y^* \\ p_{sy} S^* & 0 & p_s S^* & p_s Y^* - D_s + p_{sy} H^* \end{pmatrix}$$

where $p_c = p_{sy} + pY$, $D_y = d + d_y$ and $D_s = d + d_s$.
We evaluate the Jacobian matrix for the solutions in eqs. (0.7–0.13) and solve the equation $\det(J(H^*, X^*, Y^*, S^*) - \lambda \mathbb{I}) = 0$. The conditions of stability that satisfy that all $\lambda_i <= 0$ are the following:

- Solutions for $S^* = 0$

1. Conditions for stability, existence and non-negativity of the solution eq. (0.7) $(H^*, X^*, Y^*, S^*) = \left(\frac{g}{d}, 0, 0, 0\right)$ are:

$$R_x < \frac{d}{g} \quad R_y < \frac{d}{g} \quad \text{and} \quad R_{sy} < \frac{d}{g}$$

   The last condition adds to those in the previous model for the analogous solution. Healthy plants can escape from infection if the turnover rate d/g of healthy hosts is above the replicative success of any virus R_x, R_y or of the combination of virus and satellite, R_{sy}, when they are transmitted together.

2. Conditions for stability, existence and non-negativity of the solution eq. (0.8) $(H^*, X^*, Y^*, S^*) = \left(\frac{d+d_x}{p_x}, \frac{g}{d+d_x} - \frac{d}{p_x}, 0, 0\right)$ are:

$$R_x > \frac{d}{g}, \quad R_x > R_y \quad \text{and} \quad R_x > R_{sy}$$

   An additional condition —compared to the model without the satellite— of $R_x > R_{sy}$ is needed for invasion of X populations. One virus must overcome not only its competitor virus, but also its possible associations.

3. Conditions for stability, existence and non-negativity of the solution eq. (0.9) $(H^*, X^*, Y^*, S^*) = \left(\frac{d+d_y}{p_y}, 0, \frac{g}{d+d_y} - \frac{d}{p_y}, 0\right)$ are:

$$R_y > \frac{d}{g}, \quad R_y > R_x \quad \text{and} \quad R_y\Big(1 - \gamma\Big) > R_{sy} \tag{0.15}$$
$$\text{where} \quad \gamma = \frac{p_s}{d + d_s}\Big(\frac{g}{d + d_y} - \frac{d}{p_y}\Big)$$

   When the satellite successfully parasitizes its associate virus, S populations co-exist with Y populations (free of the satellite). However, parasitism is not unavoidable and for condition in eq. (0.15) the virus gets rid of the satellite. This equation is translated into a threshold which forces the satellite to reach reproductive ratios —only when

it coinfects with the helper virus— much higher than the reproductive success of the virus alone; otherwise, the satellite cannot be maintained in the population. In addition, this implies that associations with satellites that milden the symptoms of infection or decrease the viral —and satellite— accumulation levels, are necessarily transient because they lead to extinction.

4. Conditions for stability, existence and non-negativity of the solution eq. (0.10) $(H^*, X^*, Y^*, S^*) = \left(\frac{d+d_x}{p_x}, X^*, \frac{g}{d+d_y} - \frac{d}{p_y} - \frac{p_x}{p_y}X^*, 0\right)$ are:

$$\begin{aligned} R_x = R_y &> \frac{d}{g}, \quad R_x = R_y > R_{sy}, \\ \frac{g}{d+d_x} - \frac{d}{p_x} &> X^* > 0 \\ \frac{d+d_s}{p_s} - \frac{p_{sy}}{p_s}\frac{d+d_x}{p_x} &> Y* > 0 \quad \text{and} \\ \frac{g}{d+d_x} - \frac{d}{p_x} > X^* &> \frac{g}{d+d_x} - \frac{d}{p_x} - \frac{p_y(d+d_s) - p_{sy}(d+d_y)}{p_x p_s} \end{aligned}$$

- Solutions for $S^* > 0$

5. Conditions of stability, existence and non-negativity of the solution eq. (0.11) when $X^* = 0$ are:

$$R_y \lesseqgtr H^* \lesseqgtr R_c, \quad H^* > R_{sy} \quad \text{and} \quad H^* > R_x$$

When $H^* = \frac{d+d_x}{p_x}$ these conditions are equivalent to the conditions of stability and existence of the solution in eq. (0.8), resulting in a bistable regime where the two solutions are stable equilibrium states. Depending on the initial conditions, the system approaches one or another solution.

6. Conditions of existence and non-negativity of the solution eq. (0.12) when all the variables are different from zero (the conditions for stability cannot be analytically derived) are:

$$R_y \lesseqgtr R_x \lesseqgtr R_c, \quad \text{and} \quad R_x > R_{sy}$$

Numerically, we found that the solution of eq. (0.12) is stable when $R_y > R_x > R_c$. This region corresponds to the brown and grey areas of co-existence in **Figure 3.9**.

7. Conditions of existence, non-negativity and stability of the degenerate solution eq.(0.13) when all the variables are different from zero are:

$$R_x = R_y = R_c > \frac{d}{g}$$

0.8 Numerical simulations and algorithm implementation

The algorithms of the models explained in the Chapters 2 and 3 were implemented in C-language. The compilation of the C-scripts was done using the compiler *gcc* version 4.8.4. The scripts models can be found in the Appendix A.1.

The phylogenetic analysis were done using the Bioinformatics Toolbox of MATLAB version 7.11.0.584 (R2010b). The Bioinformatics Toolbox allows to read and write Newick-formatted tree files and translate them into MATLAB workspace phylogenetic tree objects.

This tool has as well a variety of built-in methods for the tree object such as: get property values, node names, calculate the patristic distances between pairs of leaf nodes and representation. Additionally to the build-in functions several other MATLAB scripts and functions were developed and appear in the Appendix A.2 section.

All the calculations were made in a Intel Xeon(R) CPU E5504 2.00GHz $\times$ 4 processor.

0.9 Network analysis

The modules of the bipartite network of gene sharing, consisting on 622 plant virus genomes, are calculated using the tool Infomap. The Infomap is an algorithm that detects communities in large networks, using the map equation framework (Rosvall and Bergstrom, 2008). The Infomap identify modules in the network by finding an efficiently coarse-grained description of how information flows on the network. For example, a group of nodes among which information flows quickly can be aggregated and described as a single highly connected module, and the modules are connected among each other by links which capture the channels of information flow between them. It is based on a random walk which is the proxy for the information flow. The novelty of this algorithm is the efficiency to describe a random walker on the network that makes it very fast in terms of computing time.

We download and install Infomap as it is explained in the website https://www.mapequation.org/code.html. We write the 622 virus species and the 102 gene families of the bipartite network in pajek format and we run infomap using the options: undirected network, bipartite network, and a 100 number of trials. It determines 10 statistically significant modules that are shown in **Figure 4.7** and explained in Chapter 4.

0.10 Weighted distances of not-ultrametric trees

For a specific branch x of length d_x of a not-ultrametric tree we can define two subtrees i and j that sprout from it up to the leaves. The distance of consecutive branches are d_i and d_j respectively. The weighted distance of the subtrees are w_i and w_j. Thus the weighted distance to the leaves w_x is:

$$w_x = \frac{d_x}{2} + \frac{d_i w_i + d_j w_j}{w_i + w_j} \tag{0.16}$$

For the specific cases that x is the branch on which two leaves converge, the weighted distance is:

$$w_x = \frac{d_x}{2} + \frac{d_i^2 + d_j^2}{d_i + d_j} \tag{0.17}$$

where d_i and d_j are the lengths of the leaves.

A more detailed explanation of these calculations can be found together with additional developed MATLAB functions in the Appendix A.2.

PART III

RESULTS AND DISCUSSION

CHAPTER 1

PREVALENCE AND PRESENCE OF MULTIPARTITE VIRUS IN THE VIROSPHERE

Chapter overview: An exhaustive search and integration of data from different available sources permits the elaboration of a detailed reference guide of multipartite viruses. It contains data on prevalence, families and genera, host range, number of genomic segments of all multipartite species found to date. The outstanding features of multipartite viruses are also discussed.

1.1 The virosphere in numbers

Most of our current knowledge on the prevalence of multipartite viruses is recorded on publicly available databases from the International Committee on Taxonomy of Viruses (ICTV) (Lefkowitz et al., 2015) and ViralZone (Hulo et al., 2011). **Figure 1.1** summarizes the distribution of multipartite and segmented viral species within the Virosphere, and highlights the heterogeneous distribution of genome types and hosts. About 14% of all annotated viral species have a multipartite genome, while 8% are segmented species (6% and 7% of the genera, respectively, according to data from 2018). There are 19 multipartite and 16 segmented families out of 143 viral families described. However, from 2015 to 2018, the ICTV has recorded 1284 new viral species, including 89 multipartite, 112 segmented species, and several unassigned genera had established new families, illustrating in this way a fast expanding of our knowledge about the Virosphere (Holmes, 2016; Li et al., 2015; Shi et al., 2016).

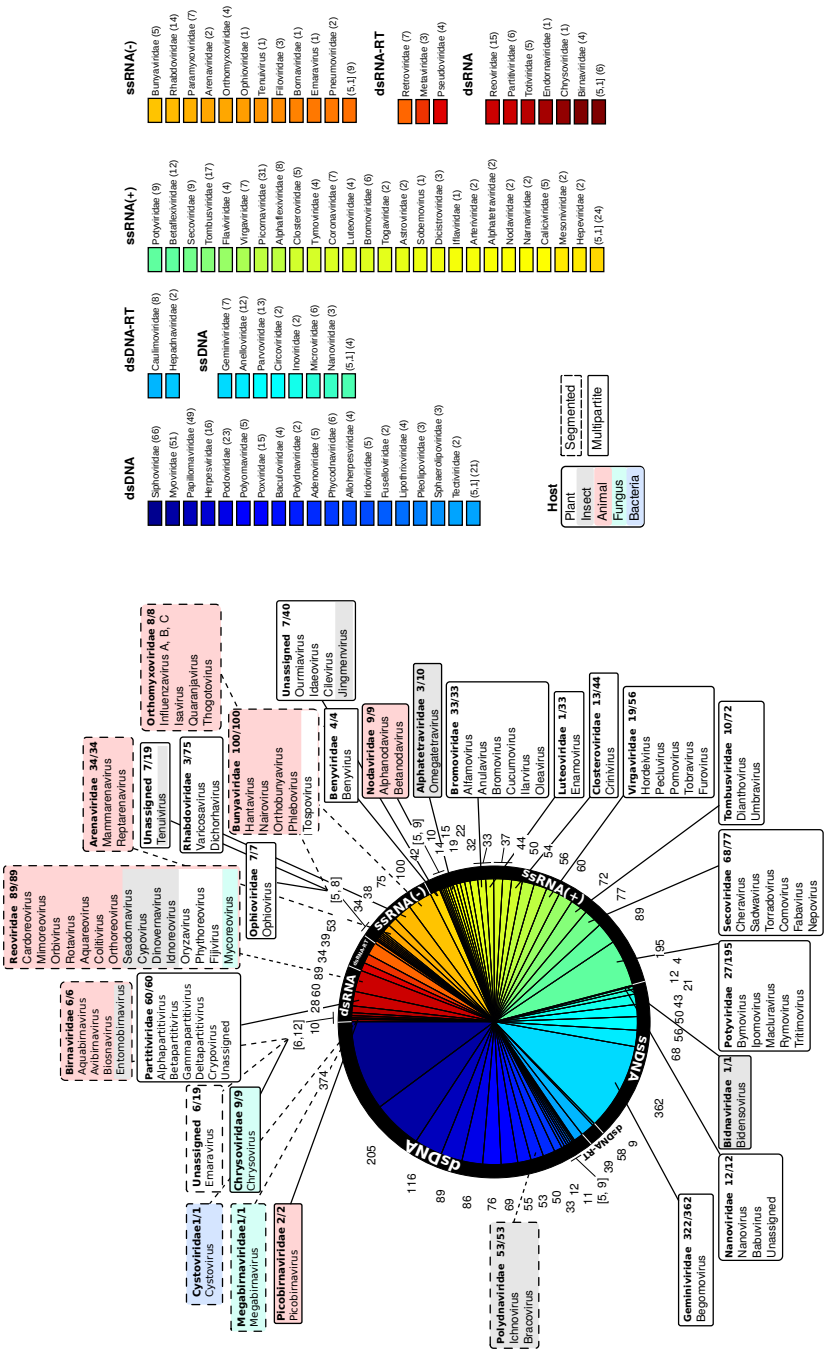
dsDNA
Siphoviridae (66)
Myoviridae (51)
Papillomaviridae (49)
Herpesviridae (16)
Podoviridae (23)
Polyomaviridae (6)
Poxviridae (15)
Baculoviridae (4)
Polydnaviridae (2)
Adenoviridae (5)
Phycodnaviridae (6)
Alloherpesviridae (4)
Iridoviridae (5)
Fuselloviridae (2)
Lipothrixviridae (4)
Pleolipoviridae (3)
Sphaerolipoviridae (3)
Tectiviridae (2)
[5,1] (21)
dsDNA-RT
Caulimoviridae (6)
Hepadnaviridae (2)
ssDNA
Geminiviridae (7)
Anelloviridae (12)
Parvoviridae (13)
Circoviridae (2)
Inoviridae (2)
Microviridae (6)
Nanoviridae (3)
[5,1] (4)
ssRNA(+)
Potyviridae (9)
Betaflexiviridae (12)
Secoviridae (9)
Tombusviridae (17)
Flaviviridae (4)
Virgaviridae (7)
Picornaviridae (31)
Alphaflexiviridae (8)
Closteroviridae (6)
Tymoviridae (4)
Coronaviridae (7)
Luteoviridae (4)
Bromoviridae (6)
Togaviridae (2)
Astroviridae (2)
Sobemovirus (1)
Dicistroviridae (3)
Iflaviridae (1)
Arteriviridae (2)
Alphatetraviridae (2)
Nodaviridae (2)
Narnaviridae (2)
Caliciviridae (5)
Mesoniviridae (2)
Hepeviridae (2)
[5,1] (24)
ssRNA(-)
Bunyaviridae (5)
Rhabdoviridae (14)
Paramyxoviridae (7)
Arenaviridae (2)
Orthomyxoviridae (4)
Ophioviridae (1)
Tenuivirus (1)
Filoviridae (3)
Bornaviridae (1)
Emaravirus (1)
Pneumovirinae (2)
[5,1] (9)
dsRNA-RT
Retroviridae (7)
Metaviridae (3)
Pseudoviridae (4)
dsRNA
Reoviridae (15)
Partitiviridae (6)
Totiviridae (5)
Endornaviridae (1)
Chrysoviridae (1)
Birnaviridae (4)
[5,1] (6)
Host
Plant
Insect
Animal
Fungus
Bacteria
Segmented
Multipartite
Reoviridae 89/89
Cardoreovirus
Mimoreovirus
Orbivirus
Rotavirus
Aquareovirus
Coltivirus
Orthoreovirus
Seadornavirus
Cypovirus
Dinovernavirus
Idnoreovirus
Oryzavirus
Phytboreovirus
Fijivirus
Mycoreovirus
Birnaviridae 6/6
Aquabirnavirus
Avibirnavirus
Blosnavirus
Entomobirnavirus
Partitiviridae 60/60
Alphapartitivirus
Betapartitivirus
Gammapartitivirus
Deltapartitivirus
Crypovirus
Unassigned
Unassigned 6/19
Emaravirus
Cystoviridae1/1
Cystovirus
Megabirnaviridae1/1
Megabirnavirus
Picobirnaviridae 2/2
Picobirnavirus
Chrysoviridae 9/9
Chrysovirus
Arenaviridae 34/34
Mammarenavirus
Reptarenavirus
Unassigned 7/19
Tenuivirus
Rhabdoviridae 3/75
Varcosavirus
Dichomavirus
Orthomyxoviridae 8/8
Influenzavirus A, B, C
Isavirus
Quaranjavirus
Thogotovirus
Bunyaviridae 100/100
Hantavirus
Nairovirus
Orthobunyavirus
Phlebovirus
Tospovirus
Ophioviridae 7/7
Ophiovirus
Unassigned 7/40
Ourmiavirus
Idaeovirus
Clevirus
Jingmenvirus
Nodaviridae 9/9
Alphanodavirus
Betanodavirus
Benyviridae 4/4
Benyvirus
Alphatetraviridae 3/10
Omegatetravirus
Bromoviridae 33/33
Alfamovirus
Anulavirus
Bromovirus
Cucumovirus
Ilarvirus
Oleavirus
Luteoviridae 1/33
Enamovirus
Closteroviridae 13/44
Crinivirus
Virgaviridae 19/56
Hordeivirus
Pecluvirus
Pomovirus
Tobravirus
Furovirus
Tombusviridae 10/72
Dianthovirus
Umbravirus
Secoviridae 68/77
Cheravirus
Sadwavirus
Torradovirus
Comovirus
Fabavirus
Nepovirus
Potyviridae 27/195
Bymovirus
Ipomovirus
Macluravirus
Rymovirus
Tritimovirus
Bidnaviridae 1/1
Bidensovirus
Nanoviridae 12/12
Nanovirus
Babuvirus
Unassigned
Geminiviridae 322/362
Begomovirus
Polydnaviridae 53/53
Ichnovirus
Bracovirus
dsRNA
ssRNA(-)
ssRNA(+)
dsDNA
dsDNA-RT
ssDNA

Figure 1.1: **Abundances of species in the Virosphere.**
Piechart showing the abundances of all currently annotated species in the ICTV (2015) for each viral family (Lefkowitz et al., 2015). Figures around the pie indicate the number of species in a given family. Colours correspond to a viral family as in the color legend, where the number of genera corresponding to each family is given in parenthesis. Families with 4 or less species are merged together. Pop charts contain multipartite (solid line) and segmented (dashed line) viral families (names in bold face), and show the number of multipartite or segmented species/total number of species. Background colors of the pop charts indicate the preferred host.
*List of families with 4 or less viral species, with the number of genera first given between brackets if different from one: *ssDNA: Bacilladnavirus 1, Spiraviridae 1, Genomoviridae 1, Bidnaviridae 1; *+RNA: Leviviridae (2) 4, Hypoviridae 4, Benyviridae 4, Ourmiavirus 3, Bacillarnavirus 3, Albetovirus 3, Sinaivirus 2, Jingmenvirus 2, Permutotetraviridae 2, Sarthroviridae 1, Carmotetraviridae 1, Barnaviridae 1, Alvernaviridae 1, Gammaflexiviridae 1, Marnaviridae 1, Roniviridae 1, Virtovirus 1, Polemovirus 1, Papanivirus 1, Labyrnavirus 1, Idaeovirus 1, Higrevirus 1, Cilevirus 1, Aumaivirus 1; *-RNA: Nyamiviridae (2) 4, Deltavirus 1, Wastrivirus 1, Crustavirus 1, Chengtivirus 1, Arlivirus 1, Anphevirus 1, Sunviridae 1, Mymonaviridae 1; *dsRNA: Amalgaviridae 4, Picobirnaviridae 2, Botybirnavirus 1, Quadriviridae 1, Megabirnaviridae 1, Cystoviridae 1.*

The viromes of eukaryotes and prokaryotes have dissimilar abundances with respect to Baltimore's classification of genome types (Baltimore, 1971). Whereas prokaryotic viruses (infecting bacteria and archaea) have a great preference for dsDNA genomes, the viruses of eukaryotes display a heterogeneous distribution among the six genome types (Koonin et al., 2015). Most viral species have a dsDNA genome, this encompassing almost all of the bacteriophages, viruses infecting archaea, and many viruses infecting eukaryotes —including animals— followed by single stranded RNA virus of positive polarity, which correspond in their vast majority to plant viruses **Figure 1.2**. In fact, only two distant families of RNA viruses infect bacteria: *Leviviridae* and *Cystoviridae* —the latter with a segmented genome and related to the animal virus family *Picorbirnaviridae* which is thought to be multipartite (Koonin et al., 2015; Krishnamurthy and Wang, 2018). Plants are infected by all types of genomes, with the exception of dsDNA, and they are the most common host followed by vertebrates —including humans. It is important to keep in mind that this picture of the Virosphere reflects in all likelihood a strong sampling bias towards those viruses with an impact in economic activities and human health **Figure 1.3**.

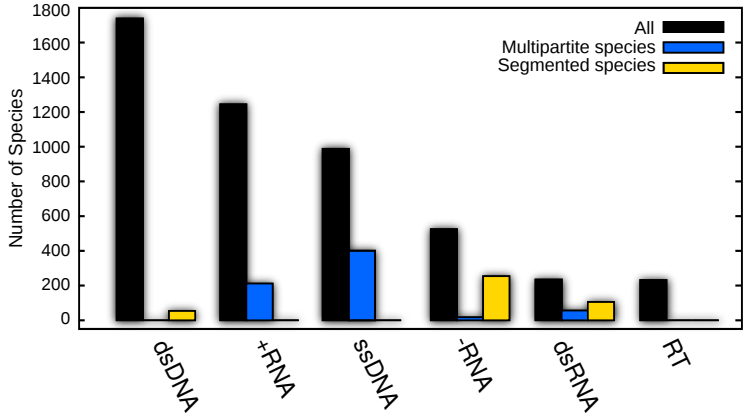

Figure 1.2: **Distribution of species according to the Baltimore classification.**
Histogram of the number of multipartite or segmented species depending on the genome type with respect to the Baltimore classification (data obtained from the ICTV (2018) (Lefkowitz et al., 2015)). Black bars correspond to all species, blue and yellow bars correspond to the abundance of multipartite and segmented species, respectively.
RT stands for reverse transcribing RNA viruses.

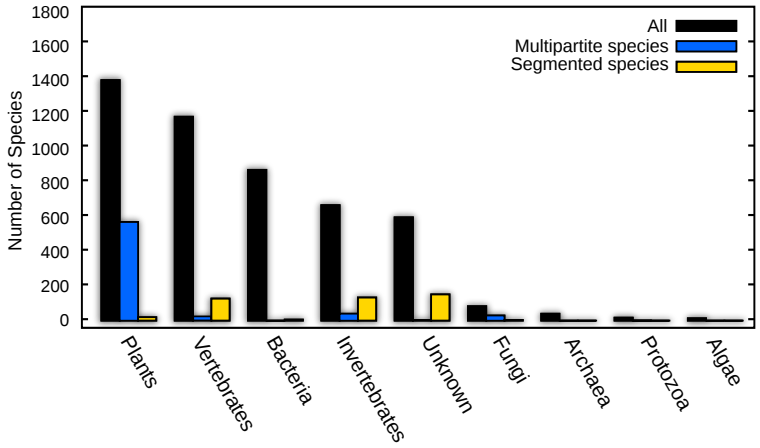

Figure 1.3: **Distribution of viral species according to host.**
Histogram of the number of species depending on the host they infect (data obtained from the publicly available database "Virus-host DB" (Mihara et al., 2016). Black bars correspond to all species, blue and yellow bars represent the abundance of multipartite and segmented species, respectively.

1.2 Prevalence and organization of segmented and multipartite viruses

Multipartite viruses infect mostly plants (90% of species and genera) followed, in much lower frequency, by invertebrates and fungi **Figure 1.3**. They are present in 14 out of 24 plant viral families, which represent 50% of all phytoviruses described. As plant pathogens, multipartite virus species display all possible genome types in phytoviruses, with the exception of retroviruses, and with the most abundant type being ssDNA genomes **Figure 1.2**. At the species level, this number is inflated due to the great success of begomoviruses, a genus which contains over 300 species; however, most families with multipartite species have positive polarity, ssRNA genomes **Table 1.1**. **Figure 1.4** represents the abundances of plant virus species per genome type and viral family, where begomoviruses are again the most abundant multipartite virus, while other families could be underrepresented at the species level due to insufficient or skewed sampling.

Most multipartite viruses (90% of species and 60% of genera) are transmitted by an invertebrate vector, like aphids, whiteflies, planthoppers, mites or nematodes, in a non-circulative, circulative or propagative manner (Power, 2000). A small amount of genera are transmitted by plant pathogens, also acting as vectors, such as fungi or protozoans (5-6%). The rest, about 35% of the genera, are vertically transmitted plant viruses. Note that vertical transmitted viruses are unaffected from the alleged cost of maintaining high MOI in which horizontally transmitted multipartite viral forms incur. Amongst 19 proposed multipartite families, 6 infect animals. *Bidnaviridae* (Hayakawa et al., 2000; Hu et al., 2016) and *Alphatetraviridae* (Tomasicchio et al., 2007) are the only families that exclusively infect animals (insects). Tenuiviruses, incorporated in 2018 to the family *Phenuiviridae* infect plants but they also replicate in their insect vectors (Falk and Tsai, 1998; Ramrez and Haenni, 1994). *Nodaviridae* is a particular multipartite family with the broadest host range which infects fishes and insects, apart from fungi and plants (Olveira et al., 2009; Selling et al., 1990). The dsRNA families *Partitiviridae* and *Picobirnaviridae* contain possible bipartite species infecting molluscs and mammals, respectively (Kim et al., 2008; Nibert et al., 2014; McDonald et al., 2016). Finally, Jingmenvirus is an unclassified genus isolated in ticks and mosquitoes (Ladner et al., 2016). The major representative infecting fungi is the family *Chrysoviridae* (Ghabrial et al., 2008) together with several genera of the family *Partitiviridae* (Nibert et al., 2014).

Table 1.1 lists all described families of multipartite viruses and their capsid structure, their abundances in terms of genera and species, the number of genomic segments they have, the hosts they infect, and their transmission mode. Looking at the numbers, there is an overwhelming majority of bipartite genera, with a rapidly dropping number of genera as the number of segments increases **Figure 1.5**. There is still a significant number of genera with 3 or 4 segments, but genera with over 4 segments are rare, with the exception of the quite diverse *Nanoviridae* family, holding up to 8 segments. It is relevant that most genera infecting hosts different from plants are bi- or tripartite, with some exceptions containing very few species. For example, *Phenuiviridae* has 4 to 6 segments and beyond plants, infect planthoppers (Falk and Tsai, 1998); jingmenviruses, which infect only insects, have 4 segments —two of them of flaviviridae origin and two other segments of unknown origin (Qin et al., 2014; Ladner et al., 2016).

Attending to genome composition, multipartite RNA plant viruses present two constitutive segments containing essential genes for infection, with extra segments with not well known functions usually needed to accomplish infection, such as in the families *Benyviridae* and *Aspiviridae*. The tripartite family *Bromoviridae* and genera *Hordeivirus* and *Pomovirus* from the family *Virgaviridae* are an exception to the former rule, since they bear

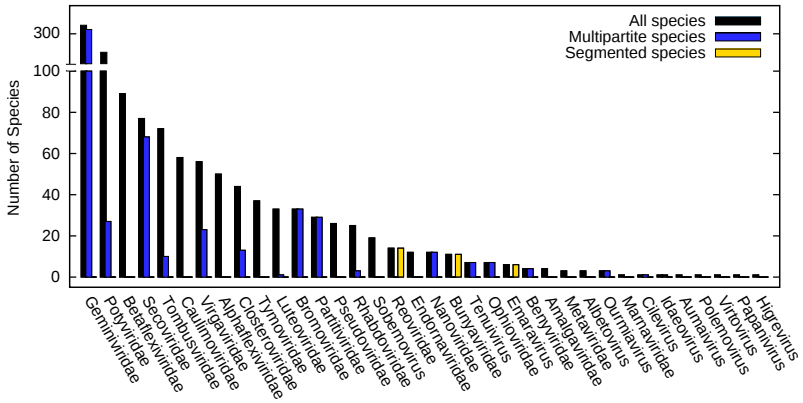

Figure 1.4: **Histogram of the number of plant virus species per viral family.** Histogram of the number of virus species per plant viral families and unassigned plant genera (data obtained from the ICTV (2015) (Lefkowitz et al., 2015)). Black bars correspond to all species, blue and yellow bars correspond to the abundance of multipartite and segmented species, respectively. Only three segmented families infect plants, *Fimoviridae (Emaravirus)*, *Peribunyniaviridae (Tospovirus)* and *Reoviridae* (4 genera).

constitutive genes distributed among all the three segments. Multipartite DNA plant viruses appear with two types of genome configurations. A representative of the first type is the genus *Begomovirus* (in the *Geminiviridae* family) which has a principal segment containing the main genes for infection. This segment is accompanied either by an auxiliary segment or by a satellite that modify host range and symptoms (Mansoor et al., 2003). The second type is represented by the *Nanoviridae* family. This is an extreme case of multipartitism, with each of the 6 to 8 separated segments coding for a gene —though not all segments are essential for infection *in vitro* (Timchenko et al., 2006).

Viruses with segmented genomes are highly represented among viruses with negative polarity, ssRNA genomes **Figure 1.2**. In particular, 13 of 15 families bearing -RNA are segmented. The complete order *Bunyavirales*, the families *Arenaviridae* and *Ortomyxoviridae* and the recently isolated families *Quiniviridae, Chuviridae* and *Yueviridae* are examples of this genome type (Maes et al., 2018; Li et al., 2015; Shi et al., 2016). In addition, 4 segmented families have dsRNA genomes, and the class of dsDNA genomes has a unique representative family, *Polydnaviridae*.

Segmented viruses infect plants, although not as efficiently as multipartite viruses do. Three segmented families are phytoviruses: *Fimoviridae (Emaravirus)*, *Peribunyniaviridae (Tospovirus)* and *Reoviridae* **Figure 1.4**. The host range is much wider in segmented viruses, as compared to multipartite viruses, and includes bacteria (*Cystoviridae*, as mentioned above), fungi, plants, and animals, with vertebrates and invertebrates being the most frequent hosts. *Reoviridae* holds the broadest host range of the segmented families. Within this family, each genera is specialized in a different host, a feature speaking for specialization that is absent in multipartite viruses.

Whereas the number of multipartite genera decreases rapidly with the number of segments, many segmented genera reach over 3 segments, suggesting different pressures are acting for each genome architecture **Figure 1.5** —the most obvious being the cost of information loss upon virus propagation, which is not affecting segmented genomes. The major-

ity are membrane-enveloped viruses, barring *Birna-*, *Megabirna-* and *Reoviridae* families **Table 1.2**. With the exception of Tenuivirus —a multipartite genus in the family of segmented virus *Phenuiviridae*— it is remarkable that viral families with segmented genomes only contain genera with this architecture, while most families with multipartite genera also contain monopartite genera. This observation suggests that a possible evolutionary pathway towards multipartite genomes is their emergence from monopartite genomes, and even also segmented genomes, as happens for tenuiviruses.

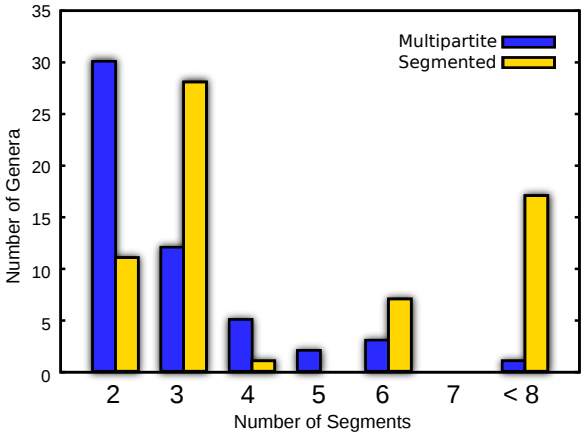

Figure 1.5: **Distribution of the number of genomic segments.**
The histogram shows the distribution of genera by the number of genomic segments they hold. The number of segments decreases rapidly for multipartite genera, a tendency that it is not followed by segmented genera (data obtained from the ICTV (2015) (Lefkowitz et al., 2015) and ViralZone (Hulo et al., 2011).

Table 1.1: Multipartite families

Capsid	Family (genera)	Species (Total)	Segments	Host	Transmission
-RNA					
	Aspiviridae (1)	7 (7)	3-4	Plants	Insects, Fungi
	Phenuiviridae (1)	7 (32)	4-6	Plants, Insects[a]	Insects (circulative[a])
	Rhabdoviridae (2)	3 (135)	2	Plants	Insects, Fungi
+RNA					
	Alphatetraviridae (1)	3 (10)	3	Insects	Oral route
	Luteoviridae (1)	2 (36)	1[b]	Plants	Insects
	Nodaviridae (2)	11 (11)	2	Vertebrates, Insects, Plants, Fungi	Contact
	Secoviridae (7)	79 (86)	2	Plants	Nematodes
	Tombusviridae (2)	12 (74)	2[b]	Plants	Contact, Insects
	Unassigned (5[c])	12	2, 2, 3, 3, 4	Plants, Insects	Pollen, Insects
	Bromoviridae (6)	36 (36)	3	Plants	Insects
	Benyviridae (1)	4 (4)	2,4-5[d]	Plants	Protozoa
	Virgaviridae (5)	20 (59)	2 or 3	Plants	Seeds, Fungi, Nematodes[e]
	Closteroviridae (1)	14 (50)	2	Plants	Insects
	Potyviridae (5)	30 (205)	2	Plants	Fungi, Insects
dsRNA					
	Chrysoviridae (1)	9 (9)	2	Fungi, Plants	Cell division
	Partitiviridae (5+15[f])	45 (60)	2	Plants, Fungi, Molluscs	Cell division
	Picobirnaviridae (1)[g]	2 (2)	2	Mammals	Oral route
ssDNA					
	Bidnaviridae (1)	1	2	Insects	Oral route
	Nanoviridae (2+1[h])	12 (12)	6 to 8 + satellites	Plants	Insects
	Geminiviridae (1)	388 (441)	2, 1+satellite	Plants	Insects

[a] *Tenuivirus* in the family *Phenuiviridae* infects plants, but is transmitted by a insect vector where it also replicates (Ramrez and Haenni, 1994).

[b] The species *Pea enantion mosaic virus* is a bipartite virus whose segments belong to two unrelated genera, *Luteovirus* and *Umbravirus* (*Tombusviridae*), each expressing their own RdRp. The same happens with *Ground rosette virus* (Syller, 2003)

[c] Unassigned genera: *Ourmiavirus* (3 species, 3 segments), *Idaeovirus* (one specie, 2 segments), *Cilevirus* (3 species, 2 segments) and *Jingmenvirus* (2 species, 4 segments) (Ladner et al., 2016).

[d] *B*enyviridae has only one genus, *Benyvirus*, which consist of 2 bipartite species and 2 species with 2 constitutive genes and a variable number of segments up to 5 of function unknown (Dall'Ara et al., 2016).

[e] *Virgaviridae* has 5 genera *Hordeivirus* and *Pomovirus* with 3 segments,and *Furovirus*, *Pecluvirus*, and *Tobravirus* with 2 segments. Only *Tobravirus* is transmitted by a nematode vector, the rest of genera are transmitted by a fungus or are seed-borne.

[f] *Partitiviridae* is a recently restructured family with 5 genera and 15 unassigned species which contains genera infecting only plants, only fungi, both simultaneously, and also one genera infecting protozoa (Nibert et al., 2014) and another infecting molluscs (Kim et al., 2008).

[g] *Picobirnaviridae* has a bipartite genome thought to be independently encapsidated (McDonald et al., 2016)

[h] *Nanoviridae* is a family with 2 genera (hexa- and octapartite) and 1 unassigned species. Many or the wild infections are accompanied by *Rep* satellites (Grigoras et al., 2008).

Table 1.2: Segmented families

Capsid	Family (genera)	Species	Segments	Host	Transmission
-RNA					
?	Qinviridae (1)	8	2	Nematodes, Arthropods	?
?	Chuviridae (1)	29	2	Arthropods, Nematodes, Vertebrates	?
?	Yueviridae (1)	2	2	Arthropods	?
	Arenaviridae (3)	41	2	Mammals, Reptiles	Zoonosis, Fomite
	Fimoviridae (1)	9	4	Plants, Chelicerata	Insects, Cell division
	Hantaviridae (4)	41	3	Mammals	Zoonosis, Oral route
	Nairoviridae (3)	16	3	Chelicerata, Mammals	Zoonosis, Insects
	Peribunyaviridae (4)	63	3	Mammals, Aves, Plants	Zoonosis, Insects
	Phasmaviridae (5)	9	3	Insects	?
	Phenuiviridae (11)	25	3	Insects, Mammals	Zoonosis, Insects
	Orthomyxoviridae (7)	9	6 or 8	Chelicerata, Mammals, Fishes	Zoonosis, Insects
dsRNA					
	Birnaviridae (4)	6	2	Molluscs, Fishes, Aves, Insects	Contact
	Megabirnaviridae (1)	1	2	Fungi	Cell division
	Quadriviridae (1)	1	4	Fungi	Cell division
	Cystoviridae (1)	7	3	Bacteria	
	Reoviridae (15)	91	8-12	Mammals, Aves, Fishes, Reptiles, Insects, Chelicerata, Plants, Fungi	Insects, Oral route, Eggs
dsDNA					
	Polydnaviridae (2)	53	10-11	Insects	Eggs

CHAPTER 2

UNDERSTANDING THE EMERGENCE OF MULTIPARTITE VIRUSES

Chapter overview: Empirical observations of multipartite viruses, the proposed advantages of multipartitism to overcome the cost of multiple infection, together with theoretical investigations of multipartite infection are reviewed in this chapter. Additionally, we focus on certain relevant empirical observations that motivate a theoretical model for the evolution of multipartitism.

2.1 Qualitative observations of multipartitism

As of today, quantitative data on multipartite viruses and their adaptive advantages is meagre, and to a large extent is restricted to the molecular properties of those viruses (Sicard et al., 2016; Dall'Ara et al., 2016; Zhang and Qu, 2015). Main empirical findings are summarized in this section and in **Table 2.1**. We do not intend to be exhaustive, rather focusing on the observations that improve the understanding of the evolutionary and adaptive mechanisms behind multipartitism.

The first (indirect) evidence for the existence of multipartite viruses arose through experiments showing a relationship between viral dose and number of local lesions steeper than predicted by the independent-action hypothesis model. This model, proposed in the 1940 decade (Price and Spencer, 1943), assumes that different viral particles do not interact during the infection process. A nice summary of this initial discovery and the additional research it triggered can be found in (Sánchez-Navarro et al., 2013). Only three families (seven genera) of multipartite viruses, all having +RNA genomes, were known in the early

1980s. The first multipartite virus with DNA genome described was a begomovirus (Haber et al., 1981), and few years later a bipartite ophiovirus with a +RNA was the first example with a filamentous nucleocapsid (Derrick et al., 1988).

A major leap in the quantitative characterization of multipartitism arrived with a cell culture experiment where the model system was an unsegmented animal virus in the Picornaviridae family, foot-and-mouth disease virus (García-Arriaza et al., 2004). After a long number of serial passages at high viral densities or multiplicities of infection (MOI), two defective and complementary viral genomes spontaneously emerged. Competition experiments between the evolved bipartite form and the wild, parental type, demonstrated the superiority of the former under high MOI conditions, while the wild parental type re-emerged through recombination as soon as the population was subjected to bottlenecks. Eventually, it was shown that defective complementary particles were more stable between infection events, this advantage sufficing to displace the parental wild type (Ojosnegros et al., 2011).

The unequal abundances of fragments qualitatively present in early experiments (French and Ahlquist, 1988; Hajimorad et al., 1991) have been unequivocally detected and quantified in other multipartite viruses (Sicard et al., 2013; Hu et al., 2016; Wu et al., 2017). Nanoviridae is a multipartite viral family with species having up to eight independent segments. Though they manage to maintain all those segments *in vivo*, it has been shown that some of the segments are actually dispensable *in vitro* (Timchenko et al., 2006). This is a puzzling observation considering the cost imposed by any additional segment on the minimum MOI for the maintenance of genetic information. The variation in orders of magnitude of the relative frequencies of the segments after an infection cycle and its dependence with the host have led to suggesting that multipartite viruses might benefit from gene-copy number regulation (Sicard et al., 2013). The situation of segment unbalance is much more demanding regarding survivability than cases where all segments are equally abundant, since the MOI must be high enough so as to guarantee that the less abundant segment is not stochastically lost (Gallet et al., 2018a).

Empirical knowledge on how multipartite viruses propagate between and within hosts is very limited. The number of viral particles transmitted from plant to plant through an insect vector–a strategy used by most plant viruses (Dietzgen et al., 2016)– ranges from 0.5 to 3.2 particles (Betancourt et al., 2008; Moury et al., 2007), although some authors suggested this number could be higher (Sicard et al., 2016). The question of how the requirement of high MOI is circumvented remains as a main open problem, strongly suggesting that mechanisms promoting non-independent propagation (Sanjuan, 2017; Taylor et al., 2014) might be at play. High MOI could be facilitated through the formation of complexes containing several heterogeneous particles, or alleviated by the fact that some vectors, like aphids, propagate in plagues, a dynamics that may result in a large number of viral particles infecting one host. Interestingly, it has been shown that a correlation between the population size of vectors and the maintenance of non-essential genomic segments exists (Betancourt et al., 2016). Once a host organism is infected, cell-to-cell propagation depends to a high extent on the molecular interactions between virus and host, and is often mediated by specific proteins (Sicard et al., 2016; Niehl and Heinlein, 2010). It has been put forward that genome parts might independently move within the tissue of an infected plant, such that the simultaneous presence of all segments in a precise cell before the infection cycle starts would not be mandatory (Sicard et al., 2016).

Table 2.1: Empirical highlights and mathematical models of multipartitism in chronollogical order

Reference	Nomenclature	Advantage of segments	Type of model	Novelty and/or empirical basis
Price and Spencer, 1943	The relationship between the dose and the number of local lesions for different plant viruses is steeper than predicted by the independent-action hypothesis model *Alfalfa mosaic virus, Tobacco necrosis virus, Tobacco ringspot virus*			
Haber et al, 1981	First description of a multipartite virus with ssDNA genome of *Bean golden mosaic virus*			
Pressing & Reaney, 1984	Multi-compartment virus	Increase in copying fidelity under high mutation rate	Quasispecies model, included thermodynamics of the (noisy) replication process	First quantitative model. Most multipartite viruses described at the time had an RNA genome
Nee, 1987	Covirus	Higher copying fidelity and replication rate	Stable states of the competition between mono-and multipartite forms. Individual selection	Weighted the opposite effects of MOI and mutations. Followed arguments in (Pressing and Reanney, 1984)
Derrick et al., 1988	First description of a ssRNA multipartite virus of negative polarity and of the virion as a filamentous nucleocapsid of *Citrus psorosis virus*			
French and Ahlquist 1988	The different accumulation levels of genomic segments for the tripartite *Brome mosaic virus* was shown			
Iltis et al., 1989	Multi-component virus	N/A	Probabilistic, included viral interference	Dynamics of infection-dilution in cell culture series. Motivated by (Price and Spencer, 1943)
Nee & Maynard Smith, 1990	Covirus, multicomponent virus	Higher copying fidelity	Deterministic, game theory	Overall and integrative review of molecular parasites. Discussion of pros and cons of multipartitism
Chao, 1991	Multi-component virus	Reduced mutational load, enhanced sex as re-assortment	Stable states of the competition between mono-and multipartite forms	Complementation and reassortment as a form of sex. Follows (Nee, 1987) and includes frequency-dependent replication
Szathmáry, 1992	Covirus, multi-compartment virus	Local clustering	Structured deme model, game theory	Local replication, effects of compartmentalization in the establishment of defective viral forms and covirus

Table continues on the next page

Nee, 2000	Covirus	Higher colonization probability	Ecological and epidemiological model	Virus-covirus coexistence is unlikely, in agreement with observations
García-Arriaza et al, 2004	An evolutionary transition from a monopartite to a fitter bipartite viral form is possible and spontaneously arises in laboratory conditions under high MOI; *Foot-and-mouth disease virus*			
Timchenko et al, 2006	Several genome segments are dispensable to develop infection in laboratory conditions, though they are maintained in vivo; *Faba bean necrotic yellow virus*			
Moury et al, 2007	Narrow bottlenecks exist during vector transmission of a plant virus. The average number of particles transmitted by an aphid is 0.5-3.2; *Potato virus Y*			
Miyashita et al, 2010	Narrow bottlenecks exist in cell-to-cell movement during tissue infection, promoting superinfection exclusion, though it does not affect genome complementation; *Soil-borne wheat mosaic virus*			
Ojosnegros et al, 2011	The advantage of the bipartite form in (García-Arriaza et al., 2004) is identified and quantified. Shorter genomes independently encapsidated are more stable. No advantages in replication rate or in the effect of mutations are detected; *Foot-and-mouth disease virus*			
Iranzo & Manrubia, 2012	Multipartite	Higher stability, lower degradation	Combinatorial stochastic model, game theory	Grounded in results by (García-Arriaza et al., 2004; Ojosnegros et al., 2011), implements frequency-dependent replication and population bottlenecks
Sicard et al, 2013	The significant imbalance in the abundances of different fragments in a multipartite virus is hypothesized to stem from a gene-copy number regulation; *Faba bean necrotic stunt virus*			
Sánchez-Navarro et al, 2013	Genome segments appear in different frequencies and lead to differences in the invasion probabilities of each particle. Infection kinetics are delayed for a tripartite virus compared to the monopartite situation (updating results in (Price and Spencer, 1943) and subsequent research); *Alfalfa mosaic virus, Nicotiana tabacum*			
Valdano et al, 2019	Multipartite virus	Higher transmissibility	Epidemic spatial model	Complex network of host contacts. Generalizes (Iranzo and Manrubia, 2012) to the ecological context
N/A: Not applicable				

In fact, it has been shown that individual segments replicate in different cells in a winner-takes-all-like strategy, with independence of the presence of products of that segments, suggesting that mechanisms of replication and expression are not necessary coupled (Sicard et al., 2019). Also, different segments might form aggregates that move as a unit among cells –a characteristic that has been described jointly with superinfection exclusion (Miyashita and Kishino, 2010; Niehl and Heinlein, 2010), while cell-to-cell propagation in the plant tissue is in general strongly affected by severe bottlenecks (Wu et al., 2017; Gopinath and Kao, 2007; Zwart et al., 2014).

Finally, the a priori narrow range of hosts of multipartite viruses has broadened with the description of jingmenviruses, able to infect mosquitoes and ticks; tentative species of this genus have been isolated in non-human primates (Ladner et al., 2016). It cannot be discarded that classifying multipartite viruses as plant viruses be a simplification resulting from a skewed and incomplete sampling (Holmes, 2016).

2.2 Proposed advantages of multipartite viruses

There are several possible advantages of multipartitism that were conjectured early in the literature. It was suggested that increased variability through segment shuffling could confer an adaptive advantage through the generation of hybrid viruses where clonal interference would be avoided (Fulton, 1980), a principle with possible applications to genetic engineering (van Vloten-Doting, 1983). Also, it was put forward that a multipartite virus could benefit from the spatial or temporal regulation in the synthesis of different proteins and genome parts (Harrison et al., 1976). The early identification of multipartitism with ssRNA genomes of positive polarity associated that particular genomic organization to genome type, and led to propose different advantages relying on RNA plasticity (Reanney, 1982). Other authors agreed that potential benefits of genome segmentation could be an increased genetic flexibility or a larger control of gene expression, and also suggested more efficient packaging and a possible increased resistance to inactivation by environmental agents as additional advantages (Lane, 1979). However, exceptions to the observations motivating most of those hypotheses soon caused their dismissal.

Two advantages entertained that caused significant controversy were the potentially lower mutational load of (shorter) segmented forms (Chao, 1991) and an increase in the replication rate of segments with respect to the monopartite parental type (Nee, 1987) in cases where this evolutionary pathway was assumed as the hypothetical origin of multipartite viral forms. Interestingly, the only experiment to our knowledge where the mutational load and the replication rate were simultaneously measured (Ojosnegros et al., 2011) could not identify any significant difference between the parental and the derived, bipartite genome.

Perhaps the variable frequency of the segments could represent a genuine advantage of multipartitism by granting additional regulation of gene expression through differences in the copy number of the genes (Sicard et al., 2013, 2019; Hu et al., 2016; Wu et al., 2017), although this possibility has not been fully tested yet.

2.3 Quantitative approaches to disclosing the advantages of multipartitism

The infection process of multipartite viruses has been mathematically addressed by several authors. It took some time to formalize the relationship between the dose D and the number of local-lesions curve and the multipartite character of the infecting virus (Fulton, 1962). Beyond dilution, the number of local lesions depends on the number of segments of the virus and on possible interferences among the infecting particles (Iltis et al., 1989). Killingtime curves measure the time required to completely kill a monolayer of culture cells as a function of the initial MOI (dose), and have different shapes depending on the virus being mono or multipartite. A formalism similar to that in (Iltis et al., 1989) has been used, together with propagation dynamics of infection by mono and bipartite viral forms to demonstrate that, in the former case, the killing time T is proportional to $\log(1/D)$, while in the latter it decreases as $1/D^2$ (Manrubia and Lázaro, 2006). That is, monopartite viruses kill the cellular monolayer qualitatively faster than any multipartite form spreading in similar conditions.

The qualitative hypotheses in the former section can be cast in the form of mathematical models that clarify their plausibility and limitations. All models introduced below assume a high MOI, needed to guarantee successful infection of the multipartite virus, and a model-dependent advantage to balance the high MOI cost. **Table 2.1** shows the main properties

of several of the mathematical models discussed in this section, highlighting the overall viewpoint and/or the employed techniques.

A broad class of models deals with the advantages enjoyed by multipartite forms compared to monopartite ones. The first mathematical model of this kind, largely inspired by Manfred Eigens theory of molecular quasispecies (Eigen, 1971) and its application to viruses (Domingo et al., 1978), was proposed by Pressing and Reanney (Pressing and Reanney, 1984). They argued that genome segmentation compensates for the high error levels affecting RNA replication, since smaller genetic subunits present a lesser target size to the various error-promoting agents. Group selection was implicit in this model, an assumption not needed if selection is to act on each individual segment. In such a scenario, the multipartite virus could displace the monopartite parental type at high mutation rates if the MOI was sufficiently high (Nee, 1987). Other models, understanding multipartite reproduction as a form of sex, proposed an alternative explanation based on the premise that co-infection groups are units of selection —but arguing that this is not group selection in the traditional sense—and that the high mutation rate is actually the selective pressure responsible for the emergence of multipartite forms (Chao, 1991). A detailed analysis of the meaning of viral sex and of effective levels of selection can be found in (Szathmáry, 1992). Viewing the problem from an ecological perspective, the stable coexistence between multipartite and monopartite cognate forms seemed highly unlikely (Nee, 2000).

A stochastic model based on the experimental results in (García-Arriaza et al., 2004; Ojosnegros et al., 2011) identified two evolutionarily stable states (Iranzo and Manrubia, 2012). For sufficiently high MOI the multipartite form displaced the parental wild type, while coexistence occurred for low MOI. The assumption that multipartite variants arise from a non-segmented parental form limits to 3 or 4 the number of segments achievable, since the MOI required to displace the parental wild type turns unrealistically high. The precise cost of MOI to maintain a multipartite genome has been specifically addressed also in a model based on demes (Szathmáry, 1992). Compartmentalization is an indirect form of group selection (Wilson, 1975) conceptually linked to game theory that was implicit in (Szathmáry, 1992; Iranzo and Manrubia, 2012). The powerful context and tools of game theory, also discussed in (Nee and Maynard-Smith, 1990; Nee, 2016) might be a fruitful avenue for research that deserves further attention.

Most mathematical approaches have used mean-field models and, in fact, the influence of space in the competition between mono and multipartite viruses remains widely unexplored. Exceptions are two structured models (Iranzo and Manrubia, 2012; Szathmáry, 1992), where mutualistic interactions occur among segments within a coinfection group, and an epidemic model developed in the context of complex network theory and based in (Iranzo and Manrubia, 2012) where it has been shown that the architecture of contacts between hosts is a key factor for the survival of multipartitism (Valdano et al., 2019). Regular contacts favour the fixation of multipartite viral forms, indirectly explaining the observed correlation between the intensification of agriculture and the radiation of most currently known plant viruses (Fargette et al., 2008; Gibbs et al., 2008; Pagan and Holmes, 2010) and, in all likelihood, also an expansion of multipartite viruses.

2.4 Emergence of multipartite viral forms through genome segmentation

There are many different evolutionary pathways to multipartitism that can be entertained. Their likelihood is not identical, as will be discussed in depth in Chapter 4. Here, we would like to discuss a simple model for the evolution of bipartite viral forms through generation

of defective, complementary variants, generated by the monopartite virus under unfaithful replication. The motivation of the model comes from a series of empirical observations that suggest that cooperation between defective mutants is a plausible evolutionary pathway towards multipartitism (García-Arriaza et al., 2004; White, 1996; Kim et al., 1997). However, there is no direct or indirect evidence of the plausibility of this pathway in natural scenarios since, as of yet, it is unclear where the different segments in a multipartite viral come form.

Any plausible scenario for the emergence of a multipartite genome organization has to ensure a high viral density if independent propagation of each segment is assumed. Otherwise, the complementation of the independent genome segments might repeatedly fail, leading to the extinction of the virus. The questions we tackle here are, what is the quantitative effect of the MOI? and, how does the mutation rate of the wild type affect its survival? The first question is directly related to the main constraint ascribed to multipartite genome for the maintenance of the genomic information; the second is inspired in quasispecies theory, where a high mutation rate might entail a cost sufficiently high so as to cause the extinction of the wild type form.

Here we investigate the effect of genome segmentation as an unavoidable mechanism for the emergence of multipartite virus in the wild. Viral infections often produce defective particles which contain incomplete genomes derived from errors in the replication process of the original virus (Bangham and Kirkwood, 1993). The deletion mutants lack some of the essential functions to complete the infection cycle by themselves, but they can be maintained in the population together with the parental virus —the wild type— through complementation. The relative abundance of these mutants in the population of viruses is constrained by the MOI (Manrubia et al., 2010), and importantly, depends on how faithful is the replication of the wild type (Manrubia et al., 2010).

In this scenario, we will analyse the joint effect of the mutation rate at which defective genomes are produced and the MOI in the evolutionary outcomes of the system. This model is based on the one introduced in (Iranzo and Manrubia, 2012), where the competition between monopartite and multipartite cognate viral forms, both present in the system from the onset and replicating without errors, was studied.

2.4.1 Model of genome segmentation

We consider an initial population of wild type virus (*wt*), that infects a population of susceptible cells at a given MOI, m. During the replication inside an infected cell, the *wt* genome produces deletion mutants with probability μ. For simplicity, we assume that only two mutant forms arise, either an N-terminal or a C-terminal fragment, and that they are mutually complementary. They act therefore as a bipartite species and can spread with variable ease depending on the multiplicity of infection, m. Populations of *wt* and the bipartite species in principle do not differ in terms of infective ability, thus they compete in equal conditions to infect new cells.

Within an infected cell c, replication of each segment is divided into two steps. First, each segment has to satisfy a complementarity condition, i.e., the cell has to be simultaneously coinfected either with the complementary segment or with the *wt* virus to guarantee the presence of all genomic functions needed to successfully complete the infection cycle. The offspring of each of the segments of the bipartite virus is proportional to the minimal amount of available complementation in the same cell, eq. (2.1). Other conditions for the number of offspring generated are possible, but no qualitative differences are expected (Iranzo and Manrubia, 2012). The second step is the linear generation of novel

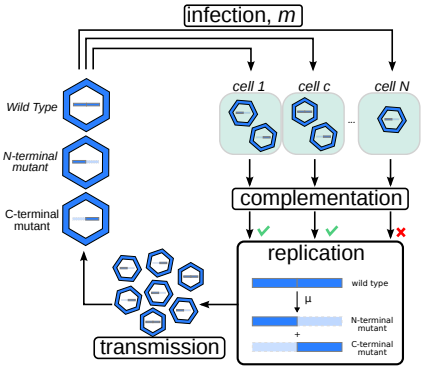

Figure 2.1: **Schematic representation of the model of genome segmentation.** Susceptible cells are infected at a multiplicity of infection m. During replication, wild type genomes produce deletion mutants N and C-terminal with probability μ. Within a cell, a given deletion mutant can complement its counterpart or be complemented by the parental wild type to succeed at replicating. The number of offspring of each of the deletion mutants is proportional to the minimal amount of available complementation counterparts in the same cell.

deletion fragments by *wt* at a probability μ in eq. (2.2). A schematics of the dynamics of the model is shown in **Figure 2.1**. The following equations describe the discrete dynamics of the replication process,

$$n_i^c = \min(n_i^c, n_{j\neq i}^c + n_{wt}^c)\,, \tag{2.1}$$

where the subindex $i = N, C$. This condition describes the number of offspring for each of the deletion mutants in each cell c.

$$\vec{n}(t=0) = \begin{pmatrix} n_{wt}(t=0) \\ 0 \\ 0 \end{pmatrix}; \quad \vec{n}(t+1) = \begin{pmatrix} (1-\mu) & 0 & 0 \\ \mu & 1 & 0 \\ \mu & 0 & 1 \end{pmatrix} \vec{n}(t) \tag{2.2}$$

where $\vec{n}(t) = (n_{wt}, n_N, n_C)$ is a vector whose components stand for the number of viral particles of each type at time t and $\vec{n}(t=0)$ is the initial condition.

We use a Poisson distribution to estimate the distribution of viruses among cells. The Poisson distribution yields the probability that a certain cell, c, gets infected by exactly k viral particles when the average number of virus particles per cell in the population is m. In this case, as three different types of viruses are infecting the cells, the probability that a cell, c, gets infected by a certain configuration of viruses $\vec{n}$, if the relative fraction of each virus are $\vec{x}$ is given by the product of three independent Poisson distributions.

This is an assumption that holds if we consider the populations are large enough such that the actual average abundances for each virus are close to mx_k, eq.(2.3).

$$Pr(\vec{n}|\vec{x}) = \prod_{k=wt,N,C} \frac{e^{-mx_k}(mx_k)^{n_k}}{n_k!} \tag{2.3}$$

We start with a population of *wt* viruses $n_{wt}(t=0)$ that infects a population of susceptible cells at an MOI, m. Once the infection cycle is completed within each cell (this entails

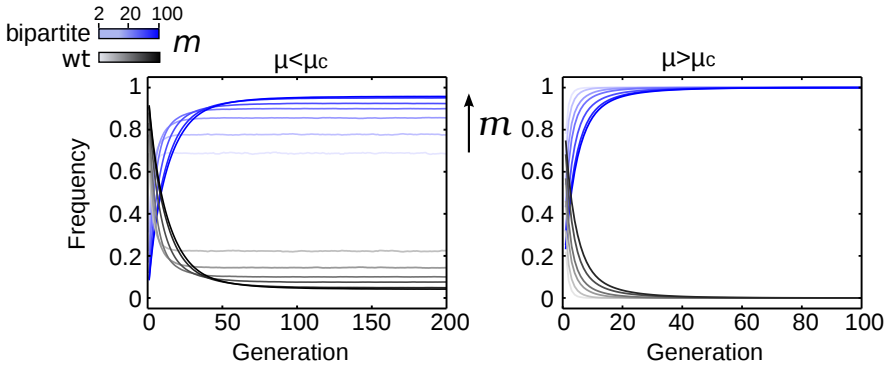

Figure 2.2: **Dynamics of virus segmentation.**
The graphics show the relative frequencies of the bipartite (blue) and *wt* (black) populations as a function of time (measured in discrete generations, or infectious cycles) for different values of m when μ is below (left panel) or above (right panel) the critical mutation rate μ_c. The color shade indicates the population dynamics for $m = [2, 100]$.

complementation-permitting replication and mutation), defective genomes have been generated for any $\mu > 0$. Then, the offspring population is added up and mixed to infect a new ensemble of susceptible cells. This process does not take into account local interactions among viruses or cells, and therefore puts the model within the class of mean-field models. The ratio between the frequencies of bipartite and *wt* viruses varies attending to the MOI and to complementation constraints. After several iterations, the population reaches an equilibrium characterized by a configuration $\vec{n}(t \to \infty) = (n^*_{wt}, n^*_N, n^*_C)$. The relative frequencies of each type are $x_i = n_i/n_{tot}$ where n_{tot} is the total amount of viral particles in the population. We measure the equilibrium composition $\vec{x}(t \to \infty) = (x^*_{wt}, x^*_N, x^*_C)$ of the viral population at different values of μ and m. For a fixed μ, the ratio x^*_N/x^*_{wt} increases as m increases, since higher multiplicities of infection facilitate complementation, as illustrated in **Figure 2.2**. In the same way, the relative frequency of bipartite virus in the population is low for low μ, and it gradually grows as μ increases until a certain critical value μ_c that causes the extinction of the *wt*, see **Figure 2.3.A**.

The critical mutation probability μ_c indicates a transition from a situation of co-existence to the extinction of the original *wt* population. Its precise value depends on m, yielding a curve in the parameter space of μ and m as is depicted in**Figure 2.3.B**. In this model, it turns out that the functional relationship between μ_c and m,

$$\mu_c = \sqrt{\frac{2}{\pi m}} \tag{2.4}$$

is equivalent to that obtained in (Iranzo and Manrubia, 2012) for the relationship between σ (which characterized the relative stability of the *wt* with respect to the defective, complementary forms, and m. In essence, it appears that any disadvantage for a *wt* virus over the bipartite form in the framework of this model promotes the fixation of the latter as m increases.

The predicted value of MOI at which the bipartite virus displaces the parental *wt* virus is relatively large even for high segmentation probabilities (high μ). Cases where the parental virus underwent segmentation into more than two fragments were explored in (Iranzo and Manrubia, 2012), and it was predicted that increasingly higher values of m were needed

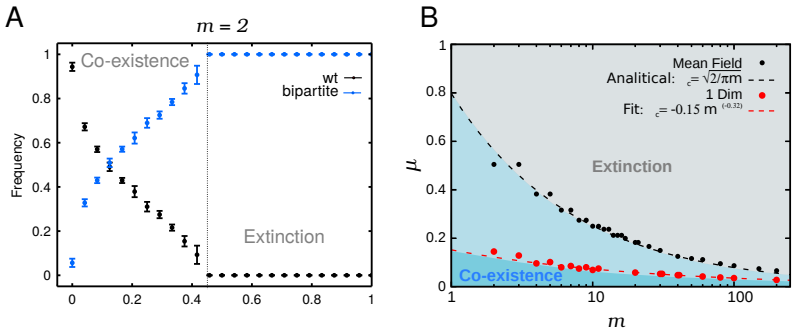

Figure 2.3: **Evolutionary regions as a function of m and μ.**
Panel A. For a fixed MOI, $m = 2$, the relative frequencies at equilibrium of the bipartite (blue) and the *wt* (black) virus change with μ. The vertical dashed line indicates the critical μ_c of the transition. This critical transition is depicted as a function of m in Panel B. Two evolutionary regions are possible: co-existence of the parental *wt* population and the defective segments (blue) and extinction of the *wt* population (grey). Black dots indicate the μ_c of the mean-field model obtained in simulations and the curve corresponds to Eq. (2.4). Red dots indicate the μ_c obtained in spatial simulations of the model, and the red dashed curve corresponds to a fit to a function of the form $\mu_c = Cm^{-\gamma}$, with fitting parameters in the figure legend and $R^2 = 0.98$.

for the *wt* population to be displaced by the multipartite virus. For example, a tetrapartite form would be fixed only for $m > 100 - 200$, well above known typical values of MOI. As the MOI detected in natural infections is very low, and it ranges from just 1 to 6 in artificial plant infections (Gutierrez et al., 2015; Gonzalez-Jara et al., 2009; Zwart et al., 2014), the likelihood of reaching, in natural conditions, values of MOI as large as those predicted by the model seems implausible. Therefore, even if genome segmentation from a parental monopartite type is a plausible evolutionary pathway leading to multipartite viruses, it does not appear as a very likely possibility that multipartite viruses with more than two segments can emerge within the scenario assumed by this model.

2.4.2 Effect of the space in genome fragmentation

The explicit introduction of space is essential when modelling the viral infection in real systems, since it may cause important qualitative changes (Cuesta et al., 2011; Boerlijst and van Ballegooijen, 2010). Susceptible hosts are often spatially structured, such as cells within the tissue or the arrangement of plants in crops fields, and so are the dynamics of infection propagation. For example, the offspring of a virus that has infected a cell within a structured tissue will be distributed preferentially among neighbouring cells, limiting the spread of the infection to the local environment. The limitations imposed by a spatial structure result in many interesting phenomena mostly driven by local clustering of viral (or pathogenic, in general) types. Local clustering is responsible for the emerging diversity in structured ecological and biological systems (Aguirre and Manrubia, 2008; Tilman and Kareiva, 1997) and can relieve the requirement of complementation in the genome segmentation model.

As an illustration of the effects of space, we consider now an array of susceptible cells in a large (one-dimensional) triangular lattice. This kind of spatial arrangement is appropriate

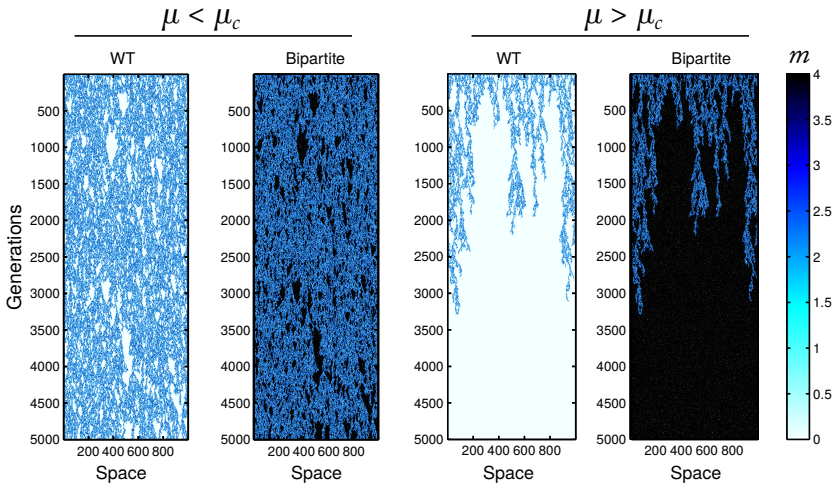

Figure 2.4: **Dynamics over time and space.**
The images show virus propagation of bipartite and *wt* populations over the generation time for a fixed MOI, $m = 4$ when μ is below (left panel) or above (right panel) the critical μ_c. A triangular lattice has been used. The color stands for m values, as shown in the color bar.

to model the propagation front of an infection (Cuesta et al., 2011). A susceptible cell c, can be infected only by viruses coming from the two closer neighbours in the upper row. The offspring of a particular cell in a row will infect, in a secondary infection, the next row: in this way, subsequent generations of viruses are tracked in space wherever the infection starts.

We have implemented this protocol of infection in the framework of the segmentation model, and looked at the relative abundances at equilibrium of each virus type as a function of μ and m. Representative examples are shown in **Figure 2.4**. For a fixed MOI, the population of *wt* and the bipartite species co-exist below a critical μ_c, generating local patches of *wt*-free virus. When the values of mutation rate μ increase beyond μ_c, we observe that the *wt* population becomes extinct after a transient whose length depends on m and on the distance to the threshold.

Similarly to the mean-field model, two evolutionary regions can be observed in the parameter space. However, the region of co-existence is much smaller compared to the mean-field model, and appears in a darker shade of blue in **Figure 2.3.B**. The change is not only qualitative, since the functional relationship between μ and m does not follow that obtained for the mean-field model; with explicit space, it can be approximated by an function of form, $\mu \propto m^{(-\gamma)}$. The bipartite viral forms can displace the parental monopartite *wt* in spatial competition under significantly less restrictive environmental conditions, as compared to the mean-field model. The effect of local clustering that emerges in an spatial model closely resembles the propagation of viral infections in plants, where it is often strongly conditioned by the two-dimensional nature of plant tissues and organization of crops fields. Eventually, this architecture might restrict the type of host where multipartite viruses emerge and thrive.

CHAPTER 3

ASSOCIATIONS IN THE VIRAL WORLD

Chapter overview: Satellites are likely remnants of genomes that make their way by coinfecting with fully fledged viruses. In return, satellites open up a range of new infection phenotypes. The ecological consequences of virus-satellite associations may compensate for the cost of coinfection and represent a stepping stone towards the emergence of a multipartite species. An original model is introduced and analysed in order to understand some ecological effects of those associations in a framework of viral competition. The implications on the evolution of multipartitism are also discussed.

There are many instances in the Virosphere of associations of viruses with kin or with subviral agents (Elena et al., 2014) that often modify the aetiology of infections (Roossinck, 2005). These associations use to be transient and imply temporary interactions between the two associates which normally operate with independence of the association. A prominent example are the so-called virus satellites that is, subviral agents that require the assistance of a specific helper virus for its replication or encapsidation (Kassanis, 1962). These associations are ubiquitous in plants, and less frequent in other hosts. Several notable exceptions infecting animals are the Hepatitis δ virus (HDV) (Makino et al., 1987), the genus *Dependoparvovirus*, in the *Parvoviridae* family (Cotmore et al., 2014; Krupovic et al., 2016) and two satellite virus infecting bees (Ribière et al., 2010), and planthoppers (Nakashima et al., 2006). Associated with the family *Totiviridae* are two isolated cases of dsRNA satellites that infect unicellular eukaryotes (Khoshnan and Alderete, 1995; Schmitt and Breining, 2002).

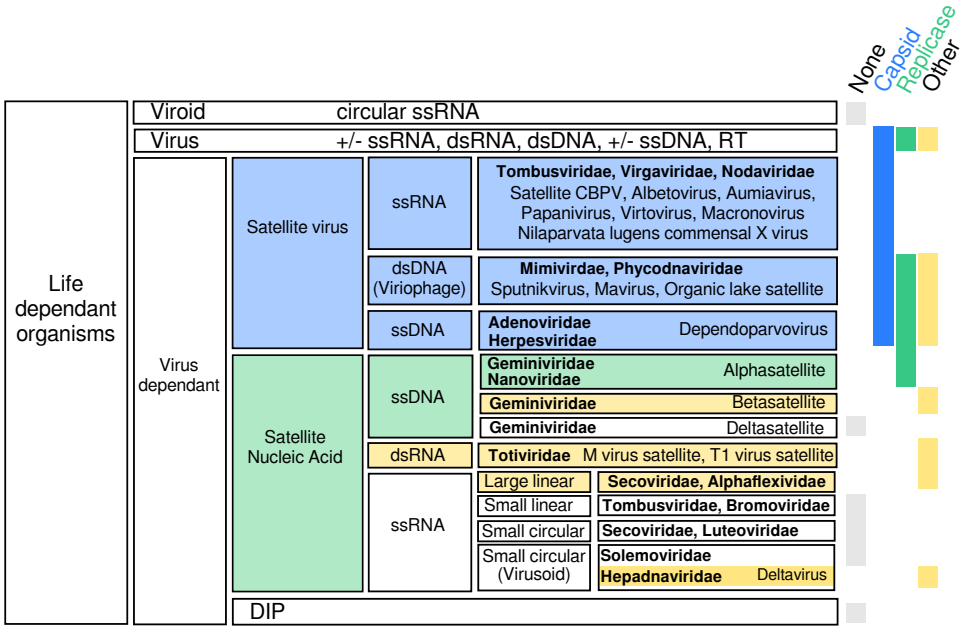

Figure 3.1: **Table of organisms depending on cellular life.**
This table shows all types of organisms/replicators/entities that depend on cellular life for their maintenance. Satellites are classified attending to the nucleic acid. Colours show the protein they encode, if any (blue: capsid, green: replicase, yellow: other, and grey: none). In each row, helper virus families are highlighted in bold font.
DIP: defective interfering particle.

A sophisticated kind of satellite-like organism are virophages. Their helper virus are typically giant viruses of the *Mimiviridae* family. Virophages normally inhibit the replication of their helper virus (Scola et al., 2008).

Another interesting class of hyper-parasites are viroid-like satellites —virusoids— consisting of a non-coding, circular ssRNA dependent on plant viruses for replication and encapsidation. Virusoids and hepatitis delta virus (HDV) are likely related to viroids (Flores et al., 2011), non-coding circular RNAs which have been exclusively described infecting plants (Symons, 1991). A detailed list of all known entities depending on cellular life can be found in **Figure 3.1**. Interestingly, some virus-satellite associations are closely linked to the presence of multipartite genomes (Murant, 1990). *Geminiviridae* is the family of plant viruses with the largest number of examples: actually, this family contains many bipartite species but also a large number of species formed by non-segmented viruses that modify their virulence and host-range through the action of a satellite (ul Rehman and Fauquet, 2009).

Mixed infections are usual for plant viruses in nature, where there is a potential for mutual interaction (Roossinck, 2005). Surprisingly, unrelated viruses within these mixtures do not seem to compete, but rather to cooperate. This observation seems in contradiction with superinfection exclusion (Gutierrez et al., 2015; Bergua et al., 2014; Gutierrez et al., 2012), where two related viruses or strains compete for (early) infection of a cell to guarantee success. The masters of synergistic interactions are potyviruses: in coinfec-

tions with a species of a different family, they usually enhance the virulence of the partner virus (Scheets, 1998; Pruss et al., 1997).

An increase of virulence and the emergence of new phenotypes may grant access to novel niches and cause persistent infections. The latter usually entail a prolonged interaction between the two associates. In this scenario, genetic recombination and gene loss are possible outcomes of the interaction, which may in this way end up in the rise of a novel species. A modular evolution, and eventually a multipartite viral species, could emerge as a consequence of independent evolutionary histories for individual genes within a genome that once belonged to two associates (Roossinck, 2005).

3.1 An intuitive classification of satellites

There is no straightforward classification of satellites, although distinguishable groups come to light whether looking at the genetic material, helper virus family, host, or encoded proteins. A currently accepted classification considers two types of satellites, as listed in **Figure 3.1**. Satellite viruses are those satellites that encode a coat protein and comprise plant satellites with jelly roll capsid proteins, dependoparvovirus and viriophages (Krupovic and Cvirkaite-Krupovic, 2012; Krupovic et al., 2016). Satellite nucleic acids are another group of satellites that may also encode a protein, but different from a capsid. For example, plant α/β–satellites and *Secoviridae* satellites encode a replicase or a replication helper protein, M virus satellite expresses a toxin and HDV encodes an antigen. In addition, this group further includes non-coding —circular— ssDNAs (Stanley et al., 1997), and RNAs with a compact folded structure that consists of a high content of base pairing —above the average (Cuesta and Manrubia, 2016)— and ribozyme activity (Hadid et al., 2017; Roossinck et al., 1992).

Another sensible way to classify satellites could consider their evolutionary origin as shown in **Figure 3.2**. It could be argued that satellites are deletion mutants formed in the same way as a defective interfering particle (DIPs), but keeping parts of the genome that still express some of the original functions or proteins. The remaining proteins of satellites are often related to an ancestral viral protein. Those satellites come from a previous virus that had reduced its genome so as to become irreversibly dependent on the parental virus. These defective interfering genomes could be identified in the presence of coinfecting viruses if they endow them with additional functions. Dependoparvoviruses, plant satellites encoding jelly roll capsid proteins, $\alpha/\beta/\gamma$-satellites and viriophages could have emerged this way; they are listed in **Figure 3.2**, upper panel (Cotmore et al., 2014; Saunders and Stanley, 1999; Krupovic et al., 2016; Krupovic and Cvirkaite-Krupovic, 2011; Fischer, 2011). Specialization of the satellite for its helper virus becomes a major issue along this process. Specialization is found to occur as one viral species associates with one to several satellite species, but not *vice versa* (Wang et al., 2017a) —with the exception of the *Dependoparvoviridae* family (Wang et al., 2017b).

An alternative hypothesis puts forward the possibility that satellites are pieces of genome "stolen" from the host that cause a fast phenotypic change of the infecting virus. This is the most plausible origin of satellite RNAs in the examples that follow, and summarized in **Figure 3.2**, bottom panel. Serial passages of a cucumovirus spontaneously generate novel satellite RNAs that derive from the host genome (Hajimorad et al., 2009; Zahid et al., 2015) and interfere with the progression of the infection (Hu et al., 2009). Tombusvirus' small RNA satellites have no similarity with their helper virus with the exception of a short initial sequence that serves as a replicase recognition site (White and Nagy, 2004).

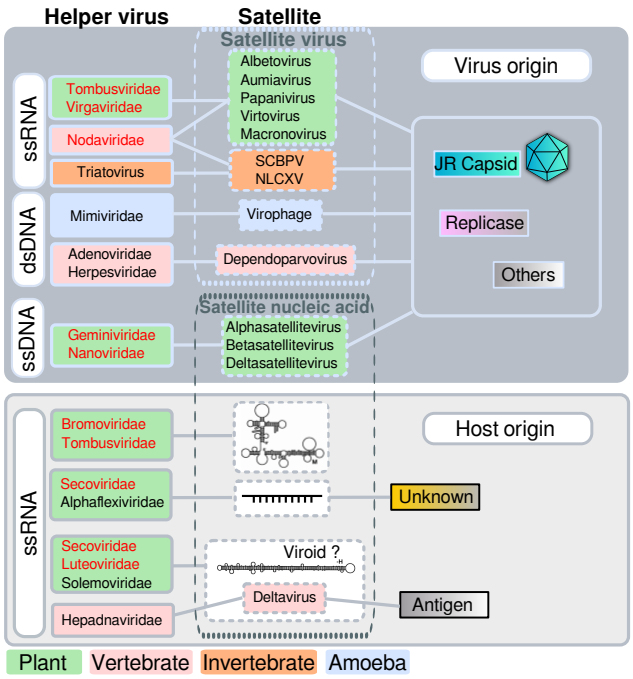

Figure 3.2: **Classification of satellites attending to their origin.**
Two groups of satellites arise attending to the evolutionary origin. Colours of the boxes determine the host. Multipartite families are in red. The satellites derived from a virus often encode one to several proteins and comprise RNA satellites encoding a jelly roll capsid protein, dependoparvovirus, virophages and $\alpha/\beta/\delta$-satellites. The other group of satellites is derived from the host —plants— and comprises the rest of RNA satellites, including viroid-like satellites.

Replicase helper proteins encoded in RNA satellites found in *Secoviridae* and *Alphaflexiviridae* infections show high resemblances between each other, but not with any other known virus (Lamprecht et al., 2013; Hadid et al., 2017). Small —circular— RNAs satellites or virusoids fold into rod-like structures that resemble those of viroids. Besides a lack of sequence similarity (Pantaleo and Burgyán, 2008; Roossinck et al., 1992; Hadid et al., 2017) these structures have shown to be energetically favoured and spontaneously appear for small, circular RNAs (Catalán et al., 2019) without significant sequence constraints. Another resemblance between viroids and virusoids is that both infect plants —with the exception of HDV— and a principle of parsimony might connect their origin to this host.

3.2 Ecological and epidemiological effects of virus associations

Viral associations usually modify the aetiology of infections because they often imply a change in the phenotype. For animal satellites such as SCBPV and NLCXV and for virophages, the association with their respective helper virus always results in a straightforward attenuation of viral symptoms (Nakashima et al., 2006; King et al., 2012; Yau et al., 2011). The exception to this rule is HDV, which induces an acute hepatitis B (Farci et al., 1988). Many more examples of satellites are known in plants, where a broader spectrum

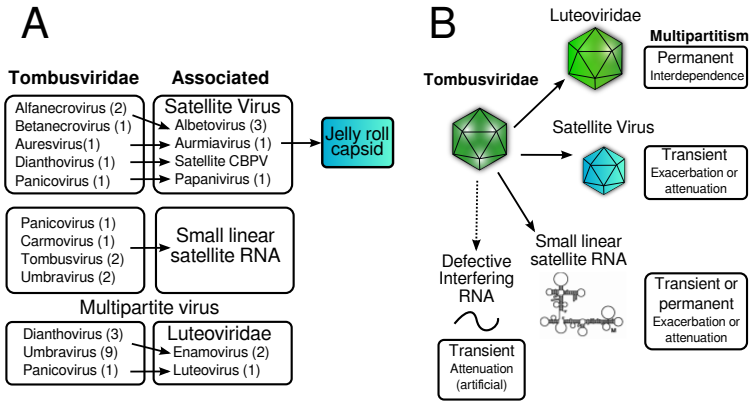

Figure 3.3: **Associations during *Tombusviridae* infections.**
Panel A shows all possible associates for the *Tombusviridae* family. The number of species is shown in parenthesis. Panel B shows the phenotypic effect of associations in A. Tombusviriuses of genus *Umbravirus* lack the capsid protein and are permanently associated with ans specie of the genus *Enamovirus* of *Luteoviridae* family, forming a bipartite species. Some umbraviruses need, beyond a luteovirus helper, the assistance of a satellite RNA to properly get transmitted by an aphid vector. However, satellite RNAs usually act transiently in tombusvirus infections, exacerbating or attenuating the symptoms of infection. Defective interfering particles have been found only in cultured conditions and not in nature, but they have shown to attenuate the symptoms of infection.

of phenotype modifications have been described. Very transient interactions may only involve an increase or decrease in viral accumulation levels. These are common effects in infections where DIPs or satellites have spontaneously emerged. Long-term interactions, in contrast, may display a variety of changes in transmission, host range, or cell tropism, ultimately leading to an interdependence —symbiosis— of the two interacting elements. In this section we are going to address the possible phenotypic changes emerging from viral associations using the family *Tombusviridae* as an example.

Tombusviridae is a family of plant virus that presents a large heterogeneity of associations, including multipartite species, and a variety of satellites. In addition, tombusviruses spontaneously generate DIPs under passaging that systematically milden the symptoms of the infection by interfering with the replication process. It has been suggested that DIPs may produce complex responses in persistent infections, apart from extinction (Manrubia et al., 2010), such as cyclical variations of viral titer and even chaotic dynamics (Kirkwood and Bangham, 1994; Bangham and Kirkwood, 1993; Moreno et al., 2017). However, tombusvirus' DIPs —and DIPs in general— are not found in the wild (Celix et al., 1997). A list of the known associations of tombusviruses is shown in **Figure 3.3.A** and an scheme of the phenotypic result of these associations in **Figure 3.3.B**.

The associations with satellites are unbalanced because their evolution is restrained from the very beginning by the evolution of the virus they parasitize. Satellite coinfections are not usually essential for the virus life cycle, and they use to be transient, with the particularity that sometimes exacerbate the symptoms of infections and others attenuate them. It has been proposed that the reason of these opposed effects is related with how the RNA satellite interacts with the immune system of the plant. If the satellite has sequence similarity with the host —host origin— it may interfere with the RNA silencing mechanism

of the plant by dismantling its immune system, ending up in an exacerbation of the infection symptoms. On the other hand, if the RNA satellite does not have sequence similarity with the host, it may trigger an acute activation of RNA silencing mechanisms, hindering the progression of the infection and leading to an attenuation of its symptoms (Hu et al., 2009). Additional examples of these phenomena are found in other plant families like *satellite Tobacco mosaic virus*, which worsens the symptoms of the helper virus (Dodds, 1998), or the tripartite virus CMV, whose association with a fourth non-essential satellite component modifies its virulence depending on the satellite component (Betancourt et al., 2011) or on the infected host (Betancourt et al., 2013). The role of satellites as virulence modifiers of their helper virus has an ecological importance in regulating both virus and host populations. Several authors explored the theoretical consequences in the ecology of the mimivirus-virophage interaction. The dynamics of virus and host population were altered considering a sequential infection —hyperparasitism— (Wodarz, 2013) or coinfection (Taylor et al., 2014). In both cases, they focused on phenotypes that emerged assuming that the virophage negatively affected virus replication, which is not the general case for satellites.

Associations with satellites have an impact beyond virus virulence, especially in those cases that imply a long-term interaction between virus and satellite. Coevolution may led to an interdependency of the two associates. For example, natural isolates of the tombusvirus *Groundnut rosette virus* (GRV), are always accompanied with an RNA satellite which is the main cause of the symptoms in groundnut and is essential for aphid transmission (Murant, 1990): without the RNA satellite, GRV is still able to infect, but it cannot be transmitted by its vector. Symbiosis has been documented for at least two tombusviruses: *Pea enation mosaic virus* (PEMV) and GRV. These viruses are unusual species of the genus *Umbravirus* which do not encode a capsid protein (Syller, 2003; Dall'Ara et al., 2016; Roossinck, 2005). They are encapsidated *in trans* by hijacking the capsid of a coinfecting virus that belongs to the family *Luteoviridae*. As pay off, they allow the entry into blocked tissues for their luteovirus partners, endowing them with a systemic movement in the plant host. The symbiosis observed between umbraviruses and enamoviruses has evolved towards the speciation of PEMV into a bipartite virus whose fragments belong to two independent genetic backgrounds. Another illustration of this phenomenon is found in the family *Geminiviridae*. Virus-satellite associations are extremely abundant in geminivirus infections. These associations became permanent in the genus *Begomovirus*: a bipartite species constituted by an ancestral geminivirus and a satellite of nanovirus origin (ul Rehman and Fauquet, 2009; Mansoor et al., 2003). Theoretical works on virus associations showed a fast transition from a mutualistic two-species to a single species dynamics, supporting the fact that mutualistic and symbiotic interactions are usually the prelude of speciation (Nee, 2000).

In order to overcome the cost of their replication, satellites should be beneficial to the helper virus, either by providing a rapid phenotypic change or an adaptive ecological response to certain environments. In a context of viral competition, an association with a satellite can quickly modify the outcome of the competition without the need of finding adaptive solutions in the form of genomic changes. Furthermore, ecological consequences of satellite associations are expected as a result of the changes in virus and host dynamics caused by the association. Altogether, we will explore theoretically in the following section the effects of a non-mandatory satellite-virus association in a context of viral competition. Our goal is to evaluate the likelihood of transient associations being a first step towards multipartitism, as it seems to have happened for PEMV and begomoviruses.

3.3 Modelling the ecological effects of a satellite in a viral competition

In this section we explore the ecological consequences of the introduction of a satellite that assists one of the two competing viruses.To this end, we devise an epidemiological model to examine the role of different parameters, in particular those related to how the interaction between virus and satellite modifies emergent phenotypes. We will first consider a model of virus competition that serves as a null model. Later, we will investigate how the introduction of a satellite that associates with one of the competing viruses modifies the dynamical outcome. Similarities and differences between both models are discussed in this section.

3.3.1 Model of viral competition

Consider a host (plant) which can be infected by two types of viruses. Healthy hosts appear at a constant rate g and decay at a rate d. The amount of susceptible hosts at time t is $H(t)$ and that of hosts infected by either virus are $X(t)$ and $Y(t)$. The model does not explicitly consider free viral populations, just susceptible or infected hosts in the different states. The model works in the mean-field approximation, and therefore assumes that hosts interact homogeneously through averaged values. Each virus is characterized by the rate $p_{x,y}$ at which it infects a susceptible host (in contact with an infected host of its class) and a parameter $d_{x,y}$ which quantifies the increase in mortality of the host due to the infected state. A scheme of the model including the parameters is depicted in **Figure 3.4**.

The equations that describe the dynamics for this system of two competing viruses are:

$$\dot{H} = g - dH - p_x XH - p_y YH \tag{3.1}$$

$$\dot{X} = p_x XH - (d + d_x)X \tag{3.2}$$

$$\dot{Y} = p_y YH - (d + d_y)Y \tag{3.3}$$

As a consequence of the symmetry of eqs (3.2) and (3.3), there will always be a virus that, in the mean-field scenario, will displace the other —with the exception of a set of points of zero-measure. For consistency, all parameters are strictly positive.

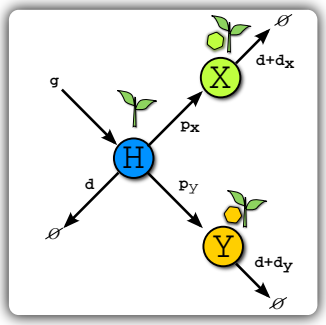

Figure 3.4: **Scheme of two virus competition.**
Healthy hosts (plants), H, are seed at a constant rate g and die with basal rate d. Plants get infected by either virus with rate p_i, being $i \in \{x, y\}$ for virus x and y. Infected plants X and Y see their basal mortality increased in an amount d_i when infected.

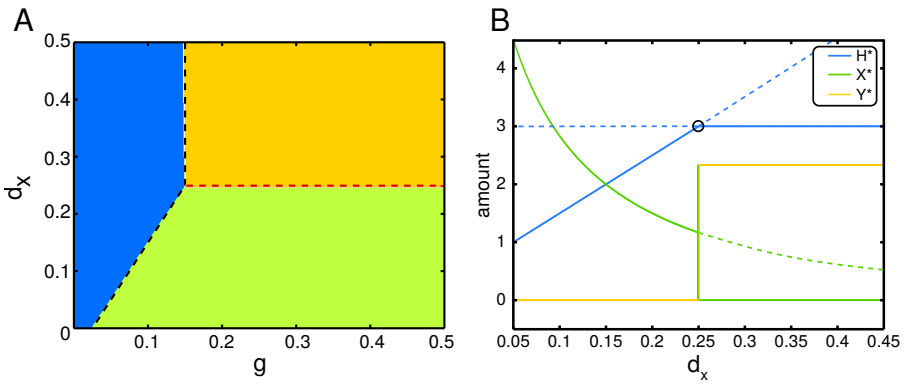

Figure 3.5: **Solutions for the model of viral competition in parameter space.** Panel A shows the solutions in the parameter space of the model of two virus competition using the parameters g and d_x. The blue region correspond to the solution where both competing viruses are extinct which satisfies the condition $d/g > R_i$. Black dashed line correspond to the condition $d/g = R_i$. The green and yellow regions correspond to solutions that satisfy $R_x > R_y$ and $R_y > R_x$ respectively, where either one of the viruses invades the population while the other becomes extinct. Red dashed line correspond to coexistence solution. Panel B shows the bifurcation diagrams of a cross section of panel A for $g = 0.5$. Dashed lines correspond to unstable solutions while solid lines correspond to stable solutions. Green and yellow regions in panel A correspond to $d_x < 0.25$ and $d_x > 0.25$. Transcritical bifurcations are distinguishable. The circle and vertical lines show the degenerate solution of coexistence of the two viruses for $d_x = 0.25$.

Results of the model of viral competition

Depending on model parameters, there are four different, positive and stable solutions: i) none of the virus is able to invade the population of hosts and both get extinct; ii) and iii) either virus x or y invades the population with the subsequent extinction of their counterpart; iv) the two virus coexist in the population. The solutions, their existence and their stabilities were obtained as explained in the Methodology section and are summarized in **Table 3.1**. The solutions of the model are shown as coloured areas in the parameter space represented in **Figure 3.5.A**, and the bifurcation maps of the solutions in **Figure 3.5.B**. The bifurcation maps show how the solutions of the system change their stability as a parameter (in this particular case, d_x) is varied. The analysis of this system presents transcritical bifurcations. In this type of bifurcations the solutions (fixed points) do not disappear after the bifurcation, instead they switch their stability. Comparable figures can be obtained with an alternative set of parameters.

In order to simplify the analysis and discussion, we define the replicative lifetime, or reproductive success, R_i, which measures the ability of either virus to invade the host population.

$$R_i = \frac{p_i}{d + d_i} \quad with \quad i \in \{x, y\} \tag{3.4}$$

The ratio d/g is a measure of the replacement or turnover time of healthy hosts. Larger d/g means longer times are needed for the appearance of susceptible healthy hosts. According to this measure, and the replicative ratios defined in eq.(3.4) we can differentiate three regions in the space of parameters for the solutions found:

Table 3.1: Conditions for stability non-negativity and existance

Eq. Fixed point (H^*, X^*, Y^*)	Conditions for existence and non negativity	Conditions of stability
(0.1) $\left(\frac{g}{d}, 0, 0\right)$	None	$R_x < d/g$ and $R_y < d/g$
(0.2) $\left(\frac{d+d_x}{p_x}, \frac{g}{d+d_x} - \frac{d}{p_x}, 0\right)$	$R_x > d/g$	$R_x > R_y$
(0.3) $\left(\frac{d+d_y}{p_y}, 0, \frac{g}{d+d_y} - \frac{d}{p_y}\right)$	$R_y > d/g$	$R_y > R_x$
(0.4) $\left(\frac{d+d_x}{p_x}, X^*, \frac{g}{d+d_y} - \frac{d}{p_y} - \frac{p_x}{p_y}X^*\right)$	$R_x = R_y > \frac{d}{g}$ and $\frac{g}{d+d_x} - \frac{d}{p_x} > X^* > 0$	None

1. **Extinction of viral populations.** If the replacement time of healthy hosts is larger than the replicative success of both viruses, the viral population goes to extinction $d/g > R_i$, $i \in \{x, y\}$. This region corresponds to the blue area in **Figure 3.5.A**. In other words, this solution corresponds to a situation in which the rate at which new susceptible hosts appear is slower than the typical time needed to invade the population.

2. **Invasion of one of the virus and extinction of the other.** Two symmetric solutions of competition appear when the reproductive ratios of the viruses are different. If either virus x or y has a replicative success higher than the one of its competitor, and also higher than the replacement time of healthy hosts ($R_i > R_j$ and $R_i > d/g$, where $i \neq j, i, j \in \{x, y\}$) the virus invades the population of hosts and its competitior gets extinct. These situations correspond to the green and yellow regions in **Figure 3.5.A**.

3. **Coexistence of both viruses.** Between the two symmetric solutions above, there is a solution of coexistence when the replicative successes of both viruses are equal and larger that the replacement time of healthy hosts, $R_y = R_x > d/g$. This solution corresponds to the red dashed line between green and yellow areas in **Figure 3.5.A**. The set of conditions for coexistence in this model has zero measure (it reduces to a line in the two-dimensional plane). Therefore, coexistence is not expected in any natural system, where slight differences between the viruses or randomness are unavoidable.

3.3.2 Model of viral competition assisted by a satellite

In the former scenario of competition between two virus, we are now introducing a possible association with a satellite. The satellite can only replicate in presence of its helper virus, which we choose to be y without loss of generality. Plants simultaneously infected by virus y and the satellite belong to a new class S, and are characterized by an increased mortality d_s. The satellite infects plants in the Y state at a rate p_s, but since it cannot replicate on its own, only plants in class Y are affected by the satellite. Simultaneous coinfection of the virus and the satellite is also possible and occurs at a rate p_{sy} under contacts between H and S plants; the former go to state S without an intermediate state Y. A coinfected plant S can infect plants in the H state only with virus y at a rate p_Y, thus contributing to plants in the Y state. A scheme of the interactions of this model can be seen **Figure 3.6**.

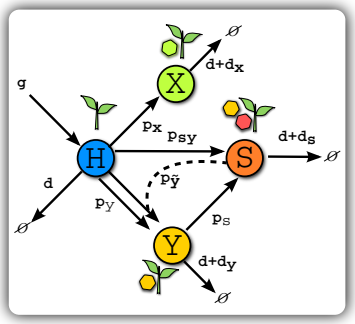

Figure 3.6: **Scheme of two virus competition with a satellite.**
Healthy hosts (plants), H, are seeded at a constant rate g and decay at a rate d. Plants get infected by either virus with a rate p_i specific of each virus, with $i \in \{x, y\}$. Infected plants X and Y increase their mortality rates in amounts d_x and d_y, respectively. Plants of type Y can get infected at a rate p_s by a satellite, thus becoming coinfected plants S with an increased mortality rate d_s. Simultaneous coinfection of H plants by the helper virus and the satellite occurs at a rate p_{sy}. A coinfected plant S can also infect H plants only with virus y at a rate p_Y.

The new set of equations reads:

$$\dot{H} = g - dH - p_x XH - p_y YH - p_{sy} SH - p_Y SH \tag{3.5}$$
$$\dot{X} = p_x XH - (d + d_x)X \tag{3.6}$$
$$\dot{Y} = p_y YH - (d + d_y)Y - p_s YS + p_Y SH \tag{3.7}$$
$$\dot{S} = p_s YS - (d + d_s)S + p_{sy} SH \tag{3.8}$$

Note that if we set $S(t = 0) = 0$ as an initial condition, this system is equivalent to the previous model. Contrary to that case, here it is not straightforward to predict which of the competing viruses or associates is going to invade the population, and depending on the new parameters different outcomes are possible. The satellite can act as a metaparasite of the helper virus and impair its propagation, in which case it might represent an advantage for the competing virus. The satellite can rescue its associate from extinction (effectively behaving as a cooperator) where the benefit is mutual. One cannot exclude situations where the satellite permits the coexistence of the previously competing viruses, given that the symmetry of the equations is now broken.

3.3.3 Results of the model of viral competition assisted by a satellite

The model has four solutions equivalent to those of the model without the satellite when the initial S population is set to 0, and three additional solutions when the satellite is considered. Two out of the three new solutions imply coexistence of all the species of the system, and the third one considers an equilibrium between Y populations, with S populations coinfected by the tandem virus-satellite, when the competitor virus x is absent. The solutions, their existence and their stabilities were obtained as explained in the Methodology section and are summarized in **Table 3.2**. In presence of the satellite, new ecological phenomena beyond competition arise for different relations of the parameters. In particular, the characteristics of the interaction between the virus and the satellite have a direct

Table 3.2: Summary of the principal solutions of the model and the conditions for stability non-negativity and existance

Eq.	**Fixed point** (H^*, X^*, Y^*, S^*)	**Conditions for existence, non negativity and stability**
(0.7)	$\left(\frac{g}{d}, 0, 0, 0\right)$	$R_x < d/g$ and $R_y < d/g$ and $R_{sy} < d/g$
(0.8)	$\left(\frac{d+d_x}{p_x}, \frac{g}{d+d_x} - \frac{d}{p_x}, 0, 0\right)$	$R_x > d/g$ and $R_x > R_y$ and $R_x > R_{sy}$
(0.9)	$\left(\frac{d+d_y}{p_y}, 0, \frac{g}{d+d_y} - \frac{d}{p_y}, 0\right)$	$R_y > d/g$ and $R_y > R_x$ and $R_y(1-\gamma) > R_{sy}$
(0.11)	$\left(H^*, 0, Y^*(H^*), S^*(H^*)\right)$	$R_y > H^* > R_c$ and $H^* > R_{sy}$ and $H^* > R_x$
(0.12)	$\left(\frac{d+d_x}{p_x}, X^*(H^*), Y^*(H^*), S^*(H^*)\right)$	$R_y > R_x > R_c$ and $R_x > R_{sy}$

impact on the outcome of the competition without an explicit modification of the helper virus infection parameters.

In order to simplify the notation, we extend our definition of reproductive success in eq. (3.4) to include the reproductive ratio of the tandem virus-satellite. We recall that the productive success R_i stands for the ability of either virus (and virus-satellite) to invade the host population.

$$R_i = \frac{p_i}{d + d_i} \quad with \quad i \in \{x, y, sy, c\} \quad where \quad p_c = p_{sy} + p_Y \tag{3.9}$$

The significance behind R_{sy} differs from that of R_c: R_{sy} is the reproductive success of the association between the virus and the satellite when they are co-transmitted in an inseparable way (simultaneous coinfection), while R_c is the replicative success of the tandem virus-satellite, despite a potential loss of the satellite during transmission. By definition, it is assured that R_c is always larger than R_{sy}, since the possibility of an independent transmission is larger than a coinfection of the two entities.

According to the parameters that define the interaction between the satellite and the helper virus, we observe three different effects on the viral competition that also satisfy conditions of existence for the solutions of the model. Therefore, the parameter space of the competition is altered under the constraints of the parameters of the satellite, as discussed below.

1. **Commensalism is neutral for the viral competition.** Coinfection with the satellite might neither benefit nor hinder the infective abilities of the helper virus. This type of parasitic interaction is called commensalism. Formally, the combination of the virus and the satellite infects with a reproductive success equal to that of the virus alone, that is $R_c = R_y$. Satisfying this condition results in no effect on the viral competition, showing that satellites acting as commensal parasites do not alter the performance of the helper virus during competition for infection. When we look at the parameter space in **Figure 3.7** we observe identical boundaries for the regions that correspond to the solutions previously described in the model without the satellite in **Figure 3.5**. Parasitism corresponds to the orange area, given by the stable solution where $Y^* \neq 0$ and $S^* \neq 0$. This solution indicates that although satellites are persistent in helper virus populations, the virus still can infect in the absence of the satellite.

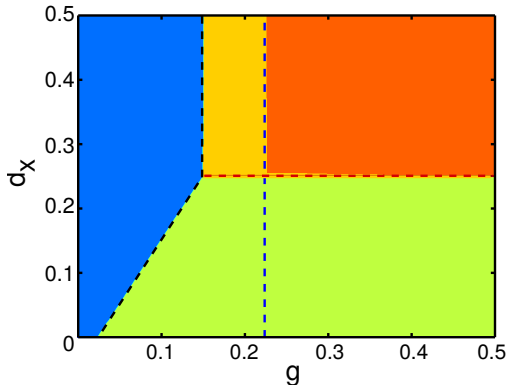

Figure 3.7: **Solutions in the parameter space for a commensal satellite.**
The graphic shows the solutions of the model of viral competition assisted by a satellite as a function of the parameters g and d_x and for the particular case $R_c = R_y$. Similarly, to the **Figure 3.5** the blue region corresponds to the infection-free situation, where both competing viruses are extinct. The green and yellow regions correspond to solutions where one of the viruses invades the population while the other becomes extinct. A novel region in orange corresponds to the coexistence of Y and S populations, with $X^* = 0$. The red dashed line corresponds to the degenerate solution of coexistence of X and Y populations. Black dashed lines correspond to the condition $g/d = R_i$ for both viruses x and y. The blue dashed line signals the boundary $g = R_y^{-1}[(d + d_s - p_{sy}/R_y)p_y/p_s + d]$.

In fact, the model shows that S populations depend on Y populations, and it is easy to demonstrate that $Y^* = 0$ implies $S^* = 0$. In addition, under certain parameters $\big(R_y(1-\gamma) > R_{sy}\big)$, where $\gamma = (g/(d+d_y) - d/p_y)p_s/(d+d_s)$, the helper virus population can get rid of the satellite. This condition is shown as a blue dashed line in **Figure 3.7**. Whether the coinfective success of the tandem virus-satellite, $R_s y$, increases, it becomes harder for the helper virus to get rid of the satellite.

2. **Mutualistic interactions can prevent the extinction of the helper virus.** Under a situation of disadvantage of the y virus in the absence of the satellite (i.e. for $R_y < R_x$, which mean extinction of the helper virus), the cooperation of the satellite can be essential to prevent its extinction. If the satellite confers a sufficiently strong increase in fitness through its association with y, the helper virus can invade the host population, leading to $Y^* \neq 0$ and $S^* \neq 0$ simultaneously. This happens when the infective abilities of the tandem virus-satellite are such that $R_c > R_x > R_y$. The outcome of this condition appears as a bistable regime in the parameter space of **Figure 3.8**, in a region that was previously only occupied by the competitor virus (pink and grey areas in **Figure 3.8 A** and **B**, respectively). A bistable regime indicates that, depending on the initial conditions, it is possible to reach one or another solution, since both are stable. In this particular case, the two possible solutions are an extinction of the helper virus —with the concomitant extinction of the associated satellite—, or an extinction of the competitor virus x. The satellite still acts as a parasite, as can be seen in the orange area in **Figure 3.8 A**, though in this case it confers an additional advantage to the helper virus by acting as a mutualistic parasite and rescue it from ex-

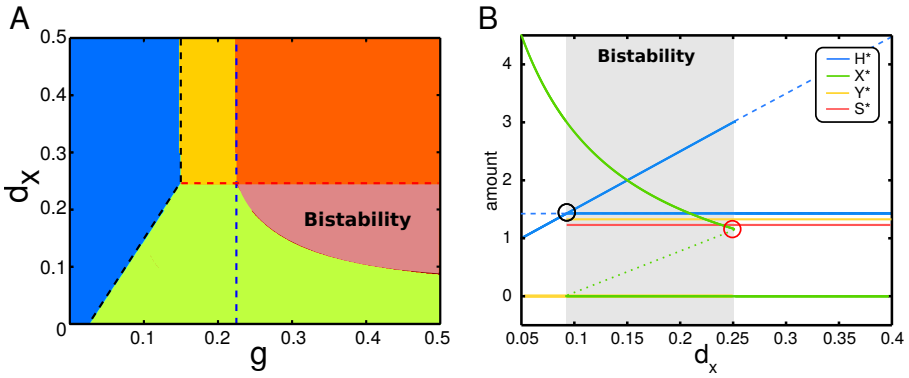

Figure 3.8: **Solutions in parameter space for a mutualistic satellite.**
Panel A. The graphic shows the solutions of the model of viral competition assisted by a satellite as a function of parameters g and d_x for the particular case $R_c < R_y$. A region of bistability is observed in pink. The regions shared with **Figure 3.7** are explained in the corresponding caption. Panel B shows the bifurcation diagrams of a cross section of panel A for $g = 0.5$. Dashed lines correspond to unstable solutions, while solid lines correspond to stable solutions. A region of bistability is shown in grey for $0.09 < d_x < 0.25$ (it corresponds to the solutions of eq. (0.8) and eq. (0.3)), meaning that, depending on the initial conditions, either populations S^* and Y^* coexist and displace X^* or *vice versa*. Transcritical bifurcations are indicated with black circles, whereas red circles correspond to the collision with the degenerated solution eq. (0.10).

tinction. Satellites which endow the helper virus with an immediate increase in fitness may be a low-cost solution to guarantee rapid adaptation to new environments.

3. **Parasitism is a burden in viral competition.** The interaction with the satellite could be expected to be a burden for the helper virus, as it has the cost of an extra replication of the satellite. Beyond that burden, parasitism can result in a clear disadvantage in a viral competition when the association with the satellite entails a loss in fitness, even if the initial situation is advantageous, i.e. $R_y > R_x$. If the replicative ratio of the combination of virus-satellite is the lowest in rank, $R_c < R_x < R_y$, then an invasion of competitor populations in a region of parameters previously forbidden (brown and grey regions in **Figure 3.9 A** and **B**, respectively) takes place. The association with a satellite that lowers the fitness of the helper virus comes as a pure disadvantage for the virus being parasitized, but endows the whole system with a region of coexistence of all species that in the previous situations did not exist. That reflects a phenomenon whereby satellites which decrease the fitness of the helper virus might allow possible coexistence, and therefore interactions, with other circulating virus in the host.

A satellite parasite can interact in three different ways with the helper virus according to the solutions of this model and the different relationships identified among the model parameters. First, when the interaction does not alter the fitness of the helper virus as it is co-transmitted with the satellite, the model predicts that satellites act as commensal parasites that, under certain conditions, the helper virus can get rid of. The model also predicts that this situation does not alter the performance of the helper virus under competition. Secondly, if a virus is fated to extinction under a situation of disadvantage in a viral

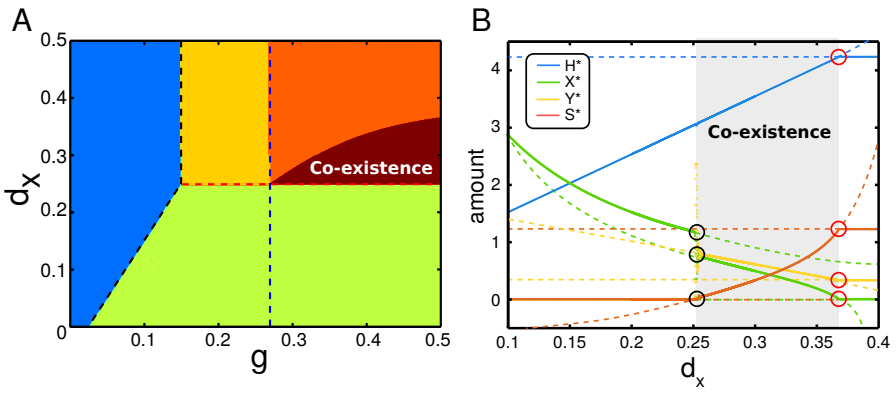

Figure 3.9: **Solutions in parameter space for the case of a parasitic satellite.** Panel A. The graphic shows the solutions of the model of viral competition assisted by a satellite as a function of parameters g and d_x, for the particular case $R_c > R_y$. A region of coexistence is observed in brown. Common regions with **Figure 3.7** are explained in the corresponding caption. Panel B shows the bifurcation diagrams of a cross-section of panel A for $g = 0.5$. Dashed lines correspond to unstable solutions, while solid lines correspond to stable solutions. A region of co-existence of all the populations is shown in grey for $0.25 > d_x > 0.37$, corresponding to the solution in eq. (0.12). Red circles correspond to transcritical bifurcations, while black circles correspond to the collision with the degenerated solution eq. (0.10).

competition, an association with a satellite that increases its fitness may rescue the virus from extinction. The association must occur when the viral density is still high enough, otherwise extinction of the virus could also happen. This type of interactions would be relevant when viruses are in their way to colonize new niches or in situations where a virus is forced to alter infections in two to several different hosts. These situations often imply that viruses are not well adapted to the new environment. Associations with satellites can temporarily increase the fitness of the virus infection, potentially allowing a fast adaptation to the new host. Third, when a virus is very virulent, an association with a satellite may milden the infective symptoms and promote coexistence of less virulent competing viruses circulating on the same host. These types of interactions are very common in virus-satellite associations, as we reviewed in the initial sections of this chapter. Coexistence in natural ecosystems open the door to interactions that may foster the genetic exchange and a variety of evolutionary outcomes.

The model predicts that associations of viruses and satellites, either promoting an increase or decrease in fitness, result in ecological phenomena that may compensate for the cost of coinfection. Coexistence and fast adaptation in competing environments are favourable outcomes of virus-satellite coinfection that may be behind the emergence of multipartite viruses.

CHAPTER 4

EVOLUTIONARY TRANSITIONS

Chapter overview: How multipartite species emerge is, as of yet, an unresolved question that may, however, have several possible answers. While the mechanisms governing the transition to multipartitism must be general and independent on the genome molecule, the possible origin of current multipartite species must be concomitant with the evolution of RNA viruses infecting eukaryotes.

4.1 On the possible origins of multipartitism

The viromes of prokaryotes and eukaryotes are qualitatively different, especially regarding the prevalence of different genome types. While prokaryotes are infected predominantly by dsDNA viruses, the diversity of RNA viruses seems to have flourished in eukaryotic cells. Together with information on the origin of some important hallmark viral proteins, a picture of a coarse-grained evolutionary hierarchy for viral genomes begins to emerge. Multipartite virus, which infect eukaryotes exclusively, are suspected to have different evolutionary origins, mainly attending to their genome type. Whereas eukaryotic RNA viruses seem to have a possible old origin in an +RNA ancestor, linked to eukaryogenesis (Koonin et al., 2015), eukaryotic ssDNA viruses could have appeared later in viral evolution, resulting from multiple events of recombination between bacterial or phytoplasma plasmids, pre-existing RNA viruses, and even viral satellites (Koonin and Dolja, 2014; Simmonds et al., 2017; Kazlauskas et al., 2017). Among ssDNA genomes, and in the Virosphere as

a whole, *Begomovirus* is the most prevalent multipartite genus, representing a main contribution to the overall abundance of multipartite species. These successful viruses use a protein called Rep to initiate replication through the rolling circle replication mechanism, which is widely used for plasmid replication in bacteria. It has been hypothesized that a recombination event between an RNA satellite and a phytoplasma plasmid could have been at the origin of Geminivirus-like viruses (Koonin and Dolja, 2014; Koonin et al., 2015). *Nanoviridae* Reps, unrelated to *Begomovirus* Reps, are more similar to those found in alpha-satellites or *Circoviruses* (Simmonds et al., 2017; Kazlauskas et al., 2017). Therefore, ssDNA viruses are unlikely to stem from a single ancestral virus.

Except for the genus *Ourmiavirus*, whose RNA-dependent RNA polymerase (RdRp) is related to the *Narnaviridae* family of RdRps (Rastgou et al., 2009), the rest of the multipartite RNA families can be rooted in one of the three major superfamilies of RNA viruses infecting eukaryotes (Picornavirus-, Alphavirus- and Flavivirus-like) (Koonin et al., 2015). These facts, together with the diversity of genomes of multipartite viruses (Figure 1.4B) suggest that multipartitism could have emerged independently a number of times in evolution.

Although the genomic "pieces" of multipartite viruses are, as far as we know, indistinguishable from those of non-segmented or segmented viruses —those pieces being as old as cellular life itself, and even older for RNA viruses—, many plant viruses seem to be recent discoveries of natural selection (Desbiez et al., 2011). Extant populations of different viral species are only up to centuries old, likely as a result of an evolutionary burst promoted through an intensification of agricultural practices (Gibbs et al., 2010). Specifically, the radiation of *Potyviridae*, *Luteoviridae* and *Sobemoviruses* can be traced back to the mid-Holocene and to the beginning of agriculture (Fargette et al., 2008; Gibbs et al., 2008; Pagan and Holmes, 2010). Also, there are no evidences of long-term co-evolution between virus and host (Desbiez et al., 2011). A recent origin of plant viruses, however, does not preclude a broad host range. A single plant virus often infects hosts across plant orders or even classes, suggesting that host switches are frequent despite a lack of obvious co-evolution (Gibbs et al., 2008). Horizontal gene transfer (HGT) and co-option of molecular functions appear in this context as a more-than-plausible mechanism for the adaptation to new hosts.

4.2 Evolutionary pathways to and from multipartitism

The different genome configurations observed in extant viruses might be solutions found after major viral families formed. In general, viruses experience frequent deletion and recombination events during replication, and HGT is common. Actually, homologies that indicate an evolutionary relationship between viruses with different genetic configurations (mono-, bi- or tripartite, in early cases) and belonging to separated taxonomical groups have been known for long (Goldbach, 1986). Gene sharing is in all likelihood directly involved in the plasticity observed in viruses at different taxonomic levels. The eventual success of a viral genome structure and configuration results from a highly contingent process. Nevertheless, viral host range does not seem to be conditioned by genome configuration: with the exception of dsDNA and retrotranscribing viruses, plants are infected by all types of genomes. There are examples of generalists such as the tripartite *Cucumber mosaic virus* (which infects over 1000 different plant hosts, both mono- and dicotyledons) or the tripartite *Tomato spotted wilt tospovirus* that infects 360 species from 50 families. However, uniform habitats and vegetative propagation may limit fitness optimization in

generalists. Strains of *Bean yellow mosaic virus* have a limited range to local cultures of domesticated plants and *Citrus tristeza virus* infections are restricted to a few genera in the *Rutaceae* (Dawson and Hilf, 1992; Moreno et al., 2008; Wylie and Jones, 2009). Anyhow, most plant viruses are generalists, and less than 10% of plant viruses infect one single host species (Power, 2008). As many as three to four different classes of viruses are often detected in an infected plant (Roossinck, 2005; Elena et al., 2014), giving plenty of opportunities to explore the joint action of different viral genomes. This behaviour could explain in part why multipartitism might have appeared repeatedly in the evolution of plant viruses. Still, possible evolutionary pathways and the specific advantage of multipartitism remain as open questions. In this section we present some ideas in this respect, following the hypothetical pathways depicted in **Figure 4.1**. Evidence to support one or another evolutionary pathway is at present uneven.

4.2.1 Transitions from non-segmented to multipartite genomes

Defective particles are routinely generated upon replication of viral genomes (Bangham and Kirkwood, 1993). Mutants with changes that preclude their viability in isolation and deletion mutants that lack genes essential to complete the viral cycle can however survive if complemented by viable genomes. Indeed, under conditions of high multiplicity of infection (MOI), defective genomes thrive thanks to the activity in *trans* of products from viable genomes. In an infection cycle, segmentation might happen through additional mechanisms. Many RNA viruses regulate gene expression through subgenomic RNAs (sgRNA). Encapsidation of sgRNA with loose or non-existent sequence signals is possible (Mandahar, 2006; Sánchez-Navarro et al., 2013). However, as defective genomes, sgRNA particles are frequently lost in presence of the wild type.

There are some *in vitro* examples of a transition from an originally non-segmented virus to a bipartite one (O'Neill et al., 1982), and some cases where genetic engineering techniques produced similar outcomes (Geigenmüller-Gnirke et al., 1991; Kim et al., 1997). These facts suggested that an evolutionary transition from a non-segmented virus to a bipartite form should be possible, given an appropriate environment. An experimental demonstration of this possibility was realized with *Foot-and-mouth disease virus* (FMDV), an animal virus that was subjected to over 200 cell culture passages at a high MOI (García-Arriaza et al., 2004). The bipartite, *in vitro* generated form that spontaneously appeared through evolution of the virus displaced the wild type in competition under the experimental conditions. Subsequent experiments demonstrated that the superiority of the bipartite form was due to an increased stability of the viral capsid, which translated into an increased particle lifespan (Ojosnegros et al., 2011). Finally, when the conditions of propagation were changed to low MOI, the two defective genomes recombined to produce a non-segmented form (Ojosnegros et al., 2011). This experiment represents a proof-of-concept that a transition to bipartitism may occur as a result of a change in the ecological context (from low to high MOI in that case).

There is some recent evidence that genome segmentation might be a rare though possible route to multipartitism in virus. This has been suggested for genus *Jingmenvirus* for which evolutionary relationships with *Flavivirus* genus (non-segmented) have been established at least for 2 out of the 4 segments of the virus —the two other segments are of unknown origin. *Flaviviruses* infect various arthropods and vertebrates, and are arthropod-borne. Several *Jingmenvirus* species have been described: *Jingmen tick virus* in ticks, mosquitoes (Qin et al., 2014) and red colobus monkey (Ladner et al., 2016), and *Guaico culex virus* in mosquitoes (Ladner et al., 2016). In addition to these multipartite species,

several other segmented flavi-like viruses have been reported infecting ticks and roundworms (Maruyama et al., 2014; Callister et al., 2008). They have been tentatively grouped into a genus, although the impossibility to culture them has made it difficult to assess their actual phylogenetic relationship. Within the species of this group, only *Guaico culex virus* has proved to be multipartite, while the rest are possibly segmented viruses. A transition to multipartitism plus a new association to other functional genes could have acted synergistically in the origin of *Jingmenvirus*.

4.2.2 Relationship between non-segmented, segmented, and multipartite viral genomes

The pathways towards and from multipartitism are largely unknown, but parsimonious mechanisms compatible with observations are much more diverse than those formalized up to now.

A first step towards multipartitism might be the generation of a segmented genome. During replication, defective genomes are unavoidably generated (S. and Baltimore, 1970), and they can persist in the population if they encapsidate in the same coat as a segmented virus **Figure 4.1.1a**, or complement each other in a favouring environment as a multipartite virus **Figure 4.1.1b**. Although in a parsimonious scenario one might expect segmented species to represent an intermediate state between non-segmented and multipartite forms, cooperation can only succeed if genome parts coincide in the same host **Figure 4.1.2**. However, mechanisms causing genome segmentation are counteracted by recombination. Defective genomes that initially compete for replication may subsequently specialize and turn into successful cooperators, such that recombinant forms cannot displace them any longer.

The transition towards multipartitism from segmented species might occur in different ways **Figure 4.1.1a-2** depending on the capsid shape. Icosahedral segmented viruses could have parsimoniously evolved towards independent encapsidation of the genome segments, a process that might underlie the origin of the families *Partitiviridae* and *Chrysoviridae*. Species within these families are vertically transmitted, a dispersion mechanism that is devoid of MOI restrictions. Their genome segments lack encapsidation signals and a mandatory co-encapsidation is not required. On the other hand, enveloped viruses whose genomes are assembled into helical nucleocapsids, such as the orders Bunya- and Mononegavirales could simply release the fragments from the enveloped membrane to become filamentous or rod-like multipartite viruses. The multipartite viruses within the families *Phenuiviridae* and *Rhabdoviridae* could have originated through this pathway. In particular, the tetrapartite tenuivirus is the only multipartite representative genus of the segmented family *Phenuiviridae*. Interestingly, the nucleocapsids of this family have homologies with filamentous capsids of bipartite genera in the families *Closteroviridae* and *Potyviridae* (Krupovic and Koonin, 2017; Kormelink et al., 2011). This transition appears as a very plausible way for a recent origin of multipartitism and it could have been facilitated through the horizontal acquisition of different or smaller capsids. A new capsid and, in general, the incorporation of novel segments, might grant access to a new collection of possible hosts, as it might happen for tenuiviruses —which actually are the only representative tetrapartite genus infecting plants of their segmented family bearing 3 segments. Similarly, the bipartite genera *Dichoravirus* and *Varicosavirus* are plant infecting viruses of the non-segmented family *Rhabdoviridae* that may be derived from insect viruses that feed from plants (Whitfield et al., 2018). The transmission from invertebrates to plants,

and therefore the adaptation to a new ecological context, might have been concomitant with multipartition (Kormelink et al., 2011).

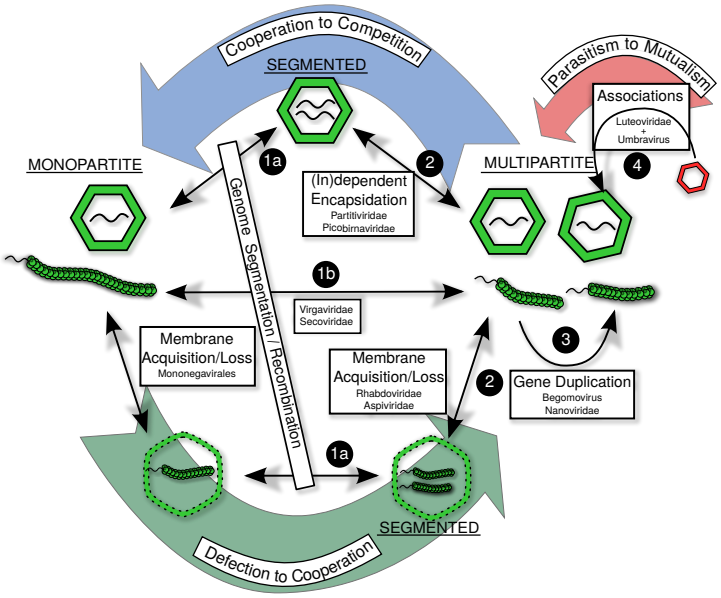

Figure 4.1: **Evolutionary pathways towards and from multipartitism.**
The chart depicts hypothetical transitions —from 1 to 4— from a non-segmented ancestor to a multipartite virus. Boxes indicate viral families or genera that could be behind these transitions. Mechanisms that could allow these transitions are annotated in arrows. (1) Segments generated from a parental non-segmented virus can in principle (1a) encapsidate in the same capsid (in the cases where the capsid is icosahedral or is enveloped by a membrane) or (1b) establish a novel multipartite species in the case of filamentous viruses. Transition (1a) could follow a pathway that first implies the acquisition of a membrane as in the case of Mononegavirales and Bunyavirales. Genome segmentation is reverted by recombination of the fragments. (2) Segmented enveloped viruses can release the fragments generating a filamentous multipartite species or, alternatively, segmented icosahedral viruses evolve towards an independent encapsidation. *De novo* segments can be originated by a gene duplication event (3) or by a virus association —especially with satellites— that becomes permanent (4).

An intermediate segmented state of the former transition might be avoided by viruses with filamentous capsids **Figure 4.1.1b**, representing an additional example of the transition from non-segmented to multipartite viruses described for *Jingmenvirus* in the previous section. Several genera in the families *Secoviridae* and *Virgaviridae* could be examples of a transition from non-segmented to multipartite species **Figure 4.1.1b** since these families comprise both types of genome organization. The origin of filamentous families like *Benyviridae* and *Virgaviridae* may include the acquisition of TMV-like capsids in order to infect plants (Krupovic and Koonin, 2017). A mechanism of co-encapsidation of genomic segments generated due to replication errors or of sgRNAs is no longer possible in the case of filamentous and rod-like capsids, so an independent —i.e. multipartite— encapsidation becomes unavoidable.

Another plausible mechanism towards multipartitism is the incorporation of novel genomic segments, be it through gene duplication or *de novo* acquisition **Figure 4.1.3-4**, as was extensively explored in the previous chapter. In the first case, several observations support the hypothesis that in a context of high viral MOI, genomic pieces could duplicate and persist in the population allowing mutations to fix and evolve to novel functions **Figure 4.1.3** representing an advantage compared to the original virus. For example, similar genome configurations are found in either of the two begomovirus fragments DNA A and B (Reiko Kikuno, 1984), and conserved regulatory regions are found in nanovirus genome fragments (Grigoras et al., 2010). *De novo* acquisition of fragments is the only mechanism towards multipartitism that may respond to ecological constraints and does not explicitly entail high MOI. The rationale behind this mechanism relies on the observation that transient associations are common in the Virosphere, specifically during co-infections. These associations —especially with satellites **Figure 4.1.4**— often modify the aetiology of infections, and therefore the ecological niche (Roossinck, 2005; Betancourt et al., 2013). The interaction between a satellite and a virus or between two different co-infecting viruses could evolve from an initial competitive or parasitic state to a cooperative or mutualistic situation. This transition could be a way to generate a novel species, as it could have happened in the origin of Pea enation mosaic virus, a hybrid of Enamovirus and Umbravirus genera, or in some Begomovirus species (Mansoor et al., 2003).

Reversibility through cooperation of the segments, and recombination of the genome segments, is likely a main force to revert to a non-segmented state. In support of this statement comes the observation that non-segmented species are also found in families containing bi- or tripartite viral genera, at odds to what is observed in segmented viral families, which do not simultaneously contain non-segmented or multipartite species.

4.2.3 Segment duplication

Replication errors and the high numbers of viable and defective genomes simultaneously found within cells might enable mechanisms analogous to gene duplication and subfunctionalization (Conant and Wolfe, 2008) in viral genomes. Complementation in *trans* opens the door to the incorporation of new mutations in defective genomes —without losing fitness— and, eventually, to the uncovering of new functions. Defective viral genomes can coexist for long in persistent infections, for instance evolving to truly hyper-parasitic forms and eventually causing the extinction of the viral population (Grande-Pérez et al., 2005). The persistence of defective segments is strongly linked to the frequency of population bottlenecks (Manrubia et al., 2010), and their presence is rarely observed *in vivo*. However, it is not unthinkable that the once defective genome might change the overall properties of the initial wild type, allow adaptation to a new ecological niche, and eventually turn out to be essential for the survival of the new, bipartite species.

Although examples of recent gene duplications in RNA viruses are not abundant, there is evidence of one such event in *Benyviridae*, a multipartite family (Simon-Loriere and Holmes, 2013). It cannot be discarded that many duplications are masked due to the rapid evolution of RNA viruses, and that remote paralogs can only be identified through the combined use of non-conventional techniques and manual curation (Kuchibhatla et al., 2014).

Gene duplications are far more frequent in viruses with DNA genomes (Shackelton and Holmes, 2004). In *Begomoviruses* (Mansoor et al., 2003; Reiko Kikuno, 1984), homologies between genes in the same segment have been identified, speaking for duplication events. A mechanism of the kind might have acted to cause the large number of segments

of *Nanoviruses*: some of the parts of this viral family are dispensable *in vitro* (Timchenko et al., 2006). Also, there is a significant degree of homology detected between some *Nanovirus* segments corresponding to regulatory sequences (Grigoras et al., 2010). Another evidence for the rapid evolution of DNA multipartitism is recombination (ul Rehman and Fauquet, 2009): the incorporation of key regulatory sequences that control the addition of foreign genes in *Begomoviruses* might be instrumental to permit the independent replication of the segments in a short time (Roberts and Stanley, 1994) and, consequently, the expansion to novel ecological niches (Lefeuvre and Moriones, 2015).

4.3 Evolution of RNA viruses

Whereas monopartite genomes do not shown a strong preference for any nucleic acid molecule in their genomes (DNA or RNA) most multipartite and segmented families are RNA viruses, as we indicated in Chapter 1. Consequently, although the mechanism by which segmented genomes originated seems to be universal and independent of the nucleic acid, the origin of segmented and multipartite viruses must be concomitant with the evolution of RNA viruses.

Understanding the evolution of RNA viruses is not possible without a comprehensive phylogeny, which is an intricate endeavour primarily due to a lack of conservation of full genes and to extensive sequence divergence. The only universal gene in this group is the virus RNA-dependent RNA polymerase (RdRp), which has suffered from strong divergence. There are, however, several conserved motifs that are required for polymerase activity. Viral RdRps are a large class of polymerases containing three catalytic domains called Thumb, Fingers and Palm —this latter exhibiting the polymerase activity— (Ferrero et al., 2018). Palm domains are a wide class of proteins also present in reverse transcriptases (RT) of retroelements and retroviruses and DNA polymreases of viral (dsDNA viruses) and cellular (DNA polymerase II) origin (Ferrero et al., 2018; Ng et al., 2008). Viral RdRps of +RNA viruses are closely related to RT of group II introns, present in prokaryotic retrotransposons, and both are thought to belong to a monophyletic group (Gladyshev and Arkhipova, 2011; Stamos et al., 2017). As we showed in Chapter 1, RNA viruses infect mostly eukaryotes, especially invertebrates and plants, with only two exceptions that infect prokaryotes: *Cystoviridae* and *Leviviridae*. However, although the origin of eukaryotic RNA viruses from prokaryotic RNA viruses remains a remote possibility (Koonin et al., 2008; Koonin, 2015), levivirus RdRps seem to be distantly related to the those of the rest of RNA viruses, and also distant from those of cystoviruses, which are more similar to eukaryotic dsRNA viruses (El Omari et al., 2013; Auguste et al., 2015).

A previous analysis of the +RNA virosphere proposed three main supergroups: picorna-like, alphavirus-like and flavivirus-like, whose ancestor was possibly an ancient picornavirus (Koonin and Dolja, 1993). In this context, the origin of -RNA viruses was anchored to an flavi-like ancestor, whereas dsRNA viruses seemed to had multiple origins in the picorna- and alphavirus-like supergroups (Koonin and Dolja, 1993; Koonin, 2015). However, the conception of RNA virus phylogeny has been recently redefined in the light of two important advances. First, new developments in virus metagenomics, combined with an extensive sampling of invertebrate taxa, have massively expanded the knowledge of the diversity of RNA viruses, especially of those groups infecting invertebrates. The expansion in RNA species eventually gave rise to at least 5 novel RNA families and permitted to fill a number of gaps in the history of RNA virus evolution (Shi et al., 2016). The corresponding phylogenetic reconstruction of the viral genomes assembled from the meta-transcriptomics

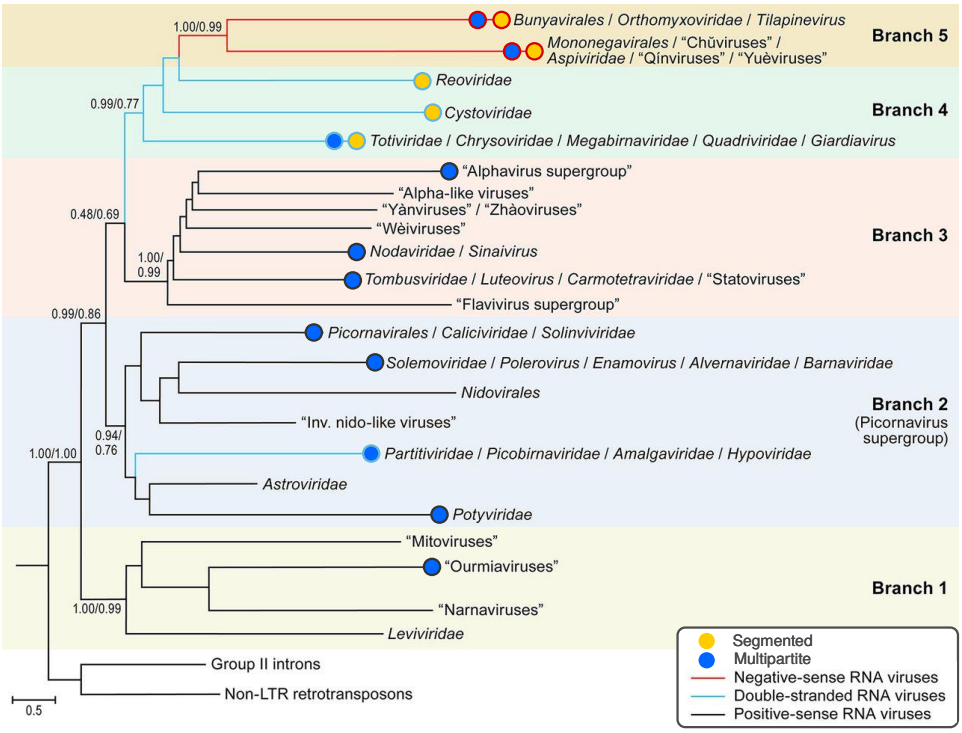

Figure 4.2: **Phylogeny of RNA virus RNA-dependent RNA polymerases and reverse transcriptases *(modified from (Wolf et al., 2018))*.**
Reconstruction of the RdRp phylogeny using a combination of methods detailed in (Wolf et al., 2018). Each branch represents a collapsed set of RdRp sequences. The 5 main branches or supergroups are obtained using supergroup alignment of several representatives within each cluster. The clusters are the result of an iterative clustering and aligning procedure where the global RdRp tree was split into separate clusters. Two independent bootstrap support values (aBayes and BOOSTER) indicated by the numerator/denominator are shown for each internal branch. LTR, long-terminal repeat. Branch colors indicate the Baltimore class, and the circles indicate segmented (yellow) and/or multipartite (blue) genome configurations that have evolved in that branch.

contemplated the unification of +RNA viruses of the three major supergroups, while -RNA viruses remained as a separate group (Li et al., 2015; Shi et al., 2016). These findings pave the way to a re-examination of RNA taxonomy with advanced techniques in deep phylogeny reconstruction (Wolf et al., 2018). In Wolf et al., a specific computational procedure was developed to coalesce 4617 RNA virus RdRps into a single phylogenetic tree with 5 major branches. The results of this work position several previously unallocated families into the RNA virus phylogeny, divide +RNA viruses in a novel way and reveal the monophyly of -RNA viruses and their apparent origin from dsRNA viruses. Also, dsRNA viruses seem to have evolved from distinct branches of +RNA viruses on at least 2 independent occasions, a result consistent with previous works (Koonin et al., 1989, 2015).

The global phylogeny of RNA viruses revealed in the paper of Wolf et al., is summarized in the phylogenetic tree of **Figure 4.2**. A brief overview of the important highlights of this work is given below.

Regarding genome molecule and configuration, Branch 1, 2 and 3 consist of +RNA viruses. Two groups of dsRNA viruses emerge from Branch 2 and 3 independently, and Branch 5 consist of all -RNA viruses. Segmented viruses are located in Branches 4 and 5, while multipartite viruses seem to have evolved independently in every branch of this phylogenetic tree. In addition, although segmented viruses have likely evolved several times in the evolution of RNA viruses —at least twice, as dsRNA and -RNA—, they are densely located into groups with multiple families. In contrast, multipartite viruses are grouped into small clusters that mostly contain only one or a few genera.

Concerning the viral taxonomy, Branch 1 consists of the families phylogenetically closer to prokaryotic introns and RT, which are prokaryotic leviviruses and their eukaryotic relatives, namely, "mitoviruses", "narnaviruses", and "ourmiaviruses" —the latter being multipartite. The quotation marks indicate that this analysis contradicts the current ICTV taxonomy, which classifies mitoviruses and narnaviruses as members of the *Narnaviridae* family, and *Ourmiavirus* as a free-floating genus. Ourmiaviruses are phytoviruses with a tripartite genome of chimeric nature that consist of 3 genes separated into 3 genetic segments. The RdRp is similar to the one found in narnaviruses, the movement protein is related to tombusviruses and the single jelly-roll capsid protein belongs to the group of picornaviruses (Rastgou et al., 2009).

Branch 2 consists of a large group of +RNA viruses infecting eukaryotes named "picornavirus supergroup". Apart from the order *Picornavirales* which includes the family *Secoviridae* with several bipartite genera, the order *Nidovirales* was moved adjacent to this group. Several families within this branch are less reliable with regard to the relative positions in the tree, such as the families *Caliciviridae, Potyviridae, Astroviridae* and *Solemoviridae* which is a natural hybrid of polerovirus and sobemovirus (Sōmera et al., 2015). Apparently, only two genera of the family *Luteoviridae* are nested within solemoviruses: *Enamovirus* —a bipartite genus— and *Polerovirus*. A lineage of dsRNA viruses constituted by partitiviruses and picobirnaviruses is located in this branch separated, from the rest of dsRNA viruses; both families are thought to be multipartite (Nibert et al., 2014; McDonald et al., 2016). Potyviruses in this branch are located next to astroviruses and show in separate linages monopartite genus (*Potyvirus*) and bipartite genera (*Macluravirus, Bymovirus* and *Ipomovirus*).

Branch 3 consists of a distinct and heterogeneous subset of +RNA viruses that coalesces the "alphavirus supergroup" along with the "flavivirus supergroup", nodaviruses, tombusviruses and some recently discovered virus groups: "statovirus", "wèivirus", "yànvirus, and "zhàovirus". A great amount of plant viruses are located in this branch. The order *Tymovirales* appears nested along with the "alphavirus supergroup" that contains a large diversity of families with multipartite genera, including the plant-infecting families *Virgaviridae, Bromoviridae, Benyviridae* and *Closteroviridae* and the family *Alphatetraviridae*, which infects insects (Tomasicchio et al., 2007). Both nodaviruses and tombusviruses contain multipartite genera, and together with the novel virus groups form a heterogeneous group in this branch. Within the "flavivirus supergroup", jingmenviruses are the only multipartite —and segmented— representatives (Ladner et al., 2016; Qin et al., 2014).

Branch 4 is anchored to Branch 3 with limited support. It consists of dsRNA viruses forming a clade separated from Branch 2 that includes totiviruses, and well known families of segmented viruses: cystoviruses, reoviruses; and a multipartite dsRNA family: *Chrysoviridae* (Ghabrial et al., 2008).

Branch 5 consists of a strongly supported lineage of all -RNA viruses considered, which is anchored very consistently to Branch 4. It splits into 2 well defined clades corresponding to two orders: *Mononegavirales* and the segmented *Bunyavirales* both containing sided groups and multipartite genera. The multipartite family *Aspiviridae*, the recently discovered segmented groups "chuvirus", "qinivirus" and "yuevirus" and the family *Rhabdoviridae* consisting of several monopartite and 2 multipartite genera: *Dichoravirus* and *Varicosavirus* also belong to the *Mononegavirales* clade. Segmented *Orthomyxovirus* together with the order *Bunyavirales* are grouped into a single clade that contains the majority of segmented families and one tetrapartite genus, *Tenuivirus*, which is the only multipartite representative of the family *Phenuiviridae*.

The global RNA phylogeny proposes a scenario where +RNA are the primary ancestors of dsRNA and -RNA viruses, which makes sense in terms of molecular logic of replication and expression strategies. It is conceivable that dsRNA evolved from +RNA viruses since it is an intermediate of +RNA replication. However, the most surprising outcome of the analysis in Wolf et al., is probably the derivation of -RNA viruses from dsRNA viruses, being the RdRp the only gene shared among these virus groups. This work provides a novel integrative picture of RNA virus evolution. Nonetheless, a better characterization of newly identified virus groups, and a structural support to the evolutionary relationships found in this work are new research directions that derive from this phylogenetic analysis (Ahola, 2019).

4.3.1 Evolution of genome configurations of RNA viruses

The evolution of the three types of RNA genomes is a main result of the work in Wolf et al. However, additional interesting information can be retrieved from the tree structure regarding the evolution of genome configuration of RNA viruses, and in particular the emergence of segmented and multipartite genomes. For example, the evolutionary rates of multipartite and segmented viruses can be extracted from the branch lengths of the tree, as well as other measurements that will be explained.

The pathways towards multipartitism presented previously and summarized in **Figure 4.1** are consistent with the findings of this analysis. It is most probable that the ancestor of multipartite viruses is a +RNA monopartite virus. Multipartite +RNA viruses can emerge directly by genome segmentation from monopartite +RNA viruses —especially in the case of filamentous viruses— such in the families clustered in Branches 1, 2, 3. A striking observation is that there are no segmented families described to date bearing +RNA genomes, with the exception of the controversial unclassified genus *Jingmenvirus* that preliminarily could contain segmented and multipartite species (Qin et al., 2014; Ladner et al., 2016). Therefore, the phylogenetic tree suggests that it is most probable that segmented viruses have flourished from animal dsRNA and -RNA viruses in at least two independent occasions (Branch 3 and 4). In addition, segmented -RNA and dsRNA viruses could be also a possible first step towards multipartitism. This appears particularly clear when looking at multipartite families with dsRNA genomes like *Chrysoviridae* or with -RNA genomes like *Tenuivirus*, both deeply nested within well known segmented groups. The mechanism governing the transition to an individual packaging of the genomic segments might depend on the capsid properties, but it could also be associated to a change in the host. Tenuiviruses and chrysoviruses infect plants, whereas most related viruses are infecting other hosts. In the case of tenuiviruses, the infection of plants entails the loss of the membrane, due the impossibility to form buds of infection in this host. Chrysoviruses could jump to plants from a fungus symbiont, and this could induce the loss of regulation

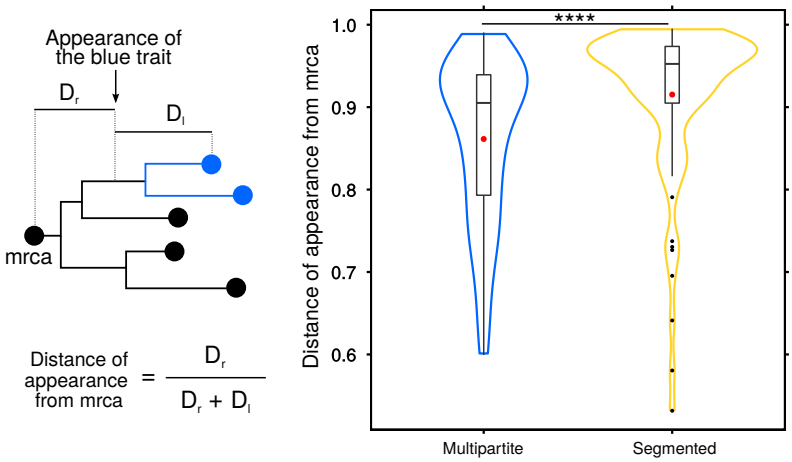

Figure 4.3: **Appearance of multipartite and segmented genomes.**
Phylogenetic distances of appearance of multipartite (blue) and segmented (yellow) genomes calculated from the global RdRp tree in (Wolf et al., 2018). The distances were calculated first identifying the specific points where multipartite and segmented traits appeared in the tree. Then, the distance to the most recent common ancestor (MRCA), D_r, and a weighted distance to the leaves, D_l, of the tree were calculated. The normalized distance of appearance of segmented and multipartite genomes was obtained and the distributions are shown in the violin plots. Statistical significance between both distributions was obtained using several two-sample tests (t-student, Wicolxon and KolmogorovSmirnov). $**** $ p-value<0.0001.

of packaging signals in the transition. Although the mechanism underlying multipartite evolution does not seem to be unique and it could depend on the specific genome configuration of the ancestor virus —whether segmented or monopartite— multipartitism appears often enough in RNA virus evolution so as to dissociate this strategy from the kind of genome molecule, which certainly does not appear as a hindrance to its origin.

4.3.2 Evolutionary distances of segmented and multipartite RNA genomes

A phylogenetic tree depicts hypothetical evolutionary relationships amongst a group of species normally based upon similarities in their physical characteristics or sequence similarity. The structure of branching reflects how species or groups evolved from a series of common ancestors. The species considered in the analysis are located in the leaves of the tree which represent the present moment. The leaves coalesce in a series of common ancestors, forming the internal nodes or branches of the tree. The branches may ultimately end in a most recent common ancestor (MRCA) in the case of rooted trees. The distance from the MRCA to a specific leaf indicates the evolutionary time or evolutionary rate of that species. The combination of methods used to build the tree of Wolf et al., results in a rooted, not-ultrametric tree, from where it is assumed that the evolutionary time that has passed for each species is different, therefore the distance from the MRCA to each leaf varies. However, despite these differences in de distances of the leaves, there is no statistically significant difference on the means of distribution of distances for species attending

to their genomic configurations. This suggests that, on average, the evolutionary rates of multipartite, segmented and monopartite RNA species are comparable.

4.3.3 Distance of appearance of segmented and multipartite RNA viruses

Multipatite and segmented species are clustered into more or less dense groups in the tree, as previously stated. Those clusters are composed by several adjacent leaves which share the same genetic configuration. We assume that the ancestor of two adjacent species which share a particular trait —or configuration— might also had that trait. Therefore, the specific node at the origin of the whole cluster is most likely the point where the trait emerged. In order to calculate when, in the evolution of RNA viruses, multipartitism and segmentation emerged, we select a cluster and travel down the tree until we find the ancestor where that specific genetic configuration appeared. The distance from this point to the root of the tree —the MRCA— is defined as D_r. The distance from the point of appearance of the trait to a specific leaf where a species within the cluster is located is defined as D_l. The differences of evolutionary time for each species in the global RNA tree, results in differences of D_l in the cluster. In order to compensate such differences we calculate a weighted distance of the cluster, as we explain in the Methodology section and is depicted in **Figure 4.3**.

We calculate the weighted distances for 346 multipartite and 368 segmented virus species included in this work. The distribution of weighted distances of appearance of segmented and multipartite genomic configurations from the MRCA are shown in **Figure 4.3**. Differences in the mean of the distributions are statistically significant, speaking for the likelihood that multipartitism appears, on the average, before segmentation in the global RNA phylogeny. That could also relate that multipartitism emerges more easily, evolutionarily speaking, than segmentation. Actually, multipartite species are spread all over the global RNA tree while segmented viruses only appear in upper Branches 4 and 5 further away from the MRCA. There are reasons to expect a shorter evolutionary time for the appearance of multipartite species as compared to segmented species. Assuming the quantitative differences observed are not a result of a biased sampling, one may argue that plants —main hosts of multipartite viruses— appeared before animals —principal hosts of segmented viruses— in the evolutionary sequence. However, whereas virus-host co-evolution may occur at the level of species (Madinda et al., 2016), there is insufficient empirical evidence correlating the evolution of eukaryotic hosts and viruses at the level of family (Geoghegan et al., 2017), with the exception of tobamoviruses in the family *Virgaviridae* (Stobbe et al., 2012; Gibbs et al., 2015). Actually, there is evidence that several plant virus groups originated from arthropod viruses (Koonin et al., 2015).

4.3.4 Diverging times for RNA species

We now address the question of the evolutionary time required for two sequences to be considered as separated species. To this end, we define the branching distance as the evolutionary distance of two sequences that have diverged. A scheme in **Figure 4.4** shows how we perform the calculation. The question of interest is whether the branching distance is affected by genome configuration, as it could be expected that multipartitism and segmentation be subjected to different evolutionary pressures.

As we did before, we identify the clusters where segmented, and/or multipartite species are grouped in the global RNA tree and we travel down the tree until we find the specific point were those configurations emerged. We extract the lengths of every branch from that point to the leaves conforming the cluster, D_b, and divide them by the total length of that

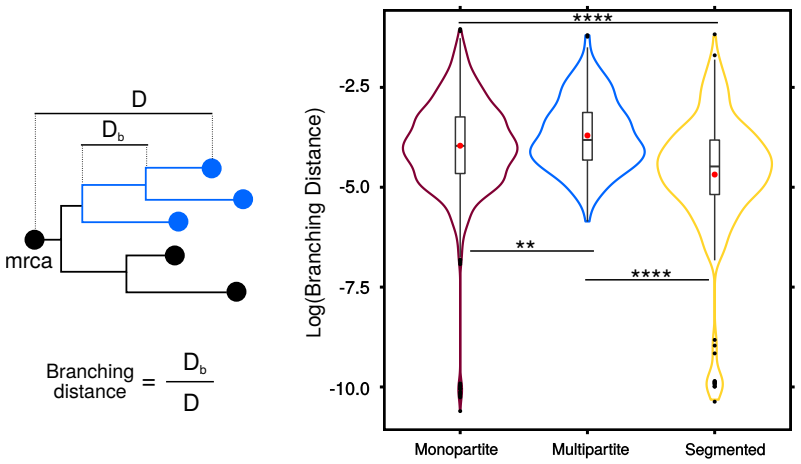

Figure 4.4: **Branching distances attending to the genome configuration.** Branching distances for monopartite (brown), multipartite (blue) and segmented (yellow) genomes were calculated from the global RdRp tree in (Wolf et al., 2018). The distances were calculated first identifying the specific points where monopartite, multipartite and segmented traits appeared in the tree. Then, the length that corresponds to the evolutionary distance of that particular trait needed to speciate is calculated, D_b. The branching distance is the ratio between D_b, and the total length from the leaves (the trait) to the most recent common ancestor (MRCA), D. $**$ p-value$<$0.01, $****$ p-value$<$0.0001.

specific branch. Since the majority of species in the tree are monopartite, the branches considered for this genome configuration are those that start at the MRCA and end in a leaf where a monopartite species is located.

Figure 4.4 shows the distributions of branching distances for the different genomic configurations. Statistically significant differences in the mean values have been found for those distributions. Segmented species have on average shorter branching distances than monopartite and multipartite species. This can be interpreted as segmented species requiring less point mutations for a sequence to diverge (*in function?*). This result could be related to segment exchange —reassortment— which is a common mechanism to generate variation in segmented virus (McDonald et al., 2016). Indeed, reassortment introduces shuffling of genome fragments that usually is accompanied by a phenotypic change that causes additional sequence variability, as compared to purely vertical inheritance (low). Overall, this result seems at odds with the hypothesis that segment shuffling might confer an advantage to multipartite species, since they seem to have, on average, larger branching distances.

4.4 Network of gene sharing of RNA viruses

If there is something that characterizes the evolution of viruses and specially RNA viruses is the high degrees of sequence divergence and an extensive HGT (Dolja and Koonin, 2018). Viruses with RNA genomes use low fidelity replicases (Drake and Holland, 1999) that cause heavy mutational loads behind the fast divergence of RNA sequences. In the last years the great advance in genomic analysis grant the access to comparative measures of

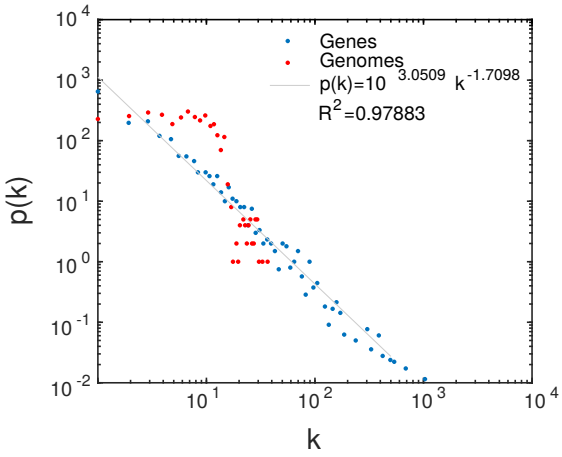

Figure 4.5: **Degree distribution of RNA network.**
The degree distributions of genes (black dots) and genomes (red dots). In the case of genes, the best fit to a power law distribution is shown.

the rates of gene gain and loss, and demonstrated that they are comparable to those of point mutation (Dolja and Koonin, 2018; Simmonds et al., 2017). HGT has been put forward as an evolutionary mechanism whose power is comparable to the fundamental, Darwinian mutation selection balance (Keeling and Palmer, 2008). Interestingly, viruses are thought to be major vehicles of action of HGT (Gilbert and Cordaux, 2017; Sano et al., 2004; Liu et al., 2010; Koonin, 2016).

High rates of sequence change and extensive gene exchange limit the applicability of traditional phylogenetic approaches to the study of virus evolution on a large time scale. In addition, in the classic phylogenetic approach, the reconstruction of the evolutionary history of a virus only takes into account the phylogeny of one gene within the genome. Single-gene phylogeny yields incongruous taxonomical trees, because genes are shared by disparate subsets of viruses: cladrograms for genes within a genome do not usually overlap. Network genomics, instead, takes into account all the evolutionary homologies within a genome, reconstructing the evolutionary history of a complete genome. The introduction of network analysis as a method that complements phylogenetic approaches revealed robust hierarchical modularity in the genomes of dsDNA viruses, bringing to light non-obvious connections among disparate groups of dsDNA viruses (Iranzo et al., 2016a,b). In such a network, each viral genome is connected to their genes by edges, and those genes are in turn linked to other genomes, forming a bipartite network structure (Iranzo et al., 2017).

The network analysis performed for the subset of RNA viruses in Wolf et al., was similar to the one in (Iranzo et al., 2016a,b). It consists in a combination of methods that begin by identifying sequence similarity among genes in the set of genomes, in order to build a list of gene families —families of homologs. This first step was performed by the group of Dr. Mart Krupovic and the specific details of the analysis are explained in (Iranzo et al., 2016b; Wolf et al., 2018). Once the families of gene homologs have been identified, the next step is to carry out an analysis of the resulting network.

In network theory, the degree of a node is the number of connections it has to other nodes. The degree distribution is therefore the probability distribution of node degrees over the whole network. A bipartite network has two types of nodes, in this case genes

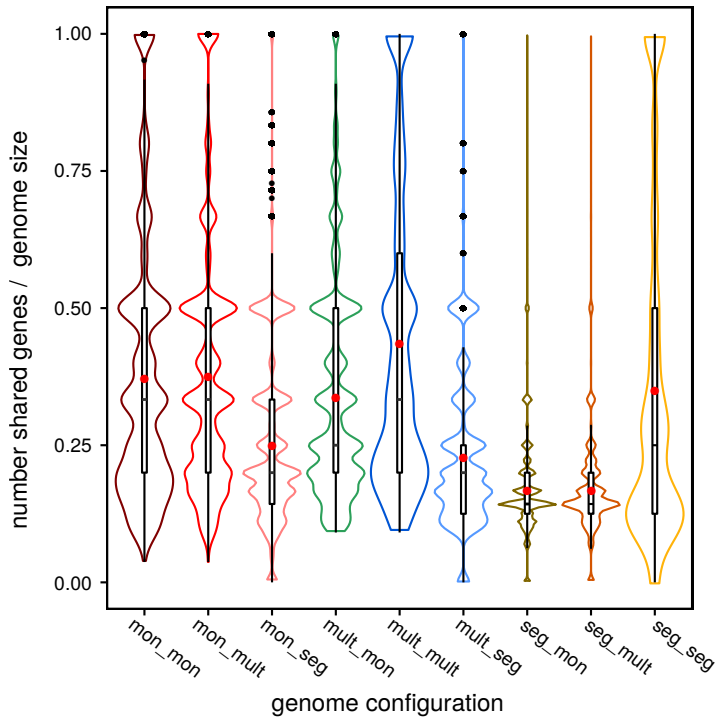

Figure 4.6: **Gene sharing distributions.**
Violin plots of gene sharing distributions attending to genome configuration. mon: monopartite, seg: segmented, mult: multipartite. Red dots represent the mean of the distributions. Boxplots represent lower and upper quartiles divided by a line with represent the median. Vertical lines are the minimum and maximum values. Black dots are outliers. In each case, the number of shared genes between a pair of genomes divided by the number of genes in the genome is shown.

and genomes. The number of genomes in which a gene family is found is represented by the degree distribution of the genes in the network. This function follows a power-law distribution $p(k) \backsim k^{-\gamma}$ shown in **Figure 4.5**. This shape means that few gene families are present in a large amount of genomes, that the average degree is not representative of "typical" nodes (there are no such nodes, actually), and that the frequency of genes with high degrees is not negligible compared to a random distribution. The degree of a genome, in contrast, follows a uniform distribution up to $k \approx 10 - 20$, which means that the fraction of genomes that are connected to one or more genes is independent of the degree. The frequency drops fast for more than approximately 10 genes, perhaps reflecting a limitation in genome length of RNA viruses (Campillo-Balderas et al., 2015).

When we look at the number of genes shared by a RNA virus species in the network we see differences attending to the genome configuration. We calculate the total number of genes shared by a genome and then we divide by the number of genes in the genome. The resulting distributions are shown in **Figure 4.6**. We observe that multipartite viruses share more genes among other multipartite species than with the rest of the genomic configurations. This result is compatible with the idea that multipartite genomes might harbour a principle of construction more plastic than the other genomic configurations. In contrast,

segmented viruses are likely to share less genes. Despite reassortment is a common feature for this genomic configuration, it seems that interspecies gene sharing is limited. In addition, low gene sharing also suggests a larger divergence of segmented families.

The degree distribution of genes and genomes in the bipartite network, and the differences in gene sharing as a function of genome configuration are empirical results that call for a mechanistic explanation. The interpretation of the preliminary results described in this section would benefit from comparisons with analogous results obtained for DNA viruses, to date only partly available. We suspect that the quantitative shape of $p(k)$ might reveal constructive processes and perhaps capture intrinsic limitations of evolutionary mechanisms in RNA versus DNA genomes, while the qualitative shape of the degree distribution (and perhaps other generic features of gene-sharing bipartite networks) should reveal generic processes in virus evolution. As for gene sharing, we believe that the significant differences between the three configurations considered might reflect how suitable each of those are for the ecological niche where they are mostly found. The application of complex network theory to genome evolution is at its beginnings and, consequently, opens more questions than yields answers. Further investigation on this and related topics seems a promising avenue to broaden our understanding of virus evolution using a different, complementary viewpoint to phylogenetic studies.

4.5 RNA plant network

Multipartite viruses infect predominantly plants. Phytoviruses usually consist of 3 main domains: a replicase, a capsid, and a movement protein. They may have auxiliary proteins for replication (helicases), translation (capping proteins), protein processing (proteases), etc. In order to explore the common features of phytoviruses, we extract from the global RNA network the subset of 622 genomes corresponding to that group. As previously done with all RNA viruses, we build the bipartite RNA plant network, which is obviously enriched in multipartite species. The resulting network is connected, i.e. does not have isolated groups. Many families are linked to few gene domains that appear as hubs in the network, and few families are connected to taxon-specific hallmark genes. The power-law distribution of gene sharing is conserved in this network and is quantitatively comparable (has similar fitting parameters) to the global RNA network shown in **Figure 4.5**.

To gain evolutionary insight we performed a modularity analysis using the tool infomap (Rosvall and Bergstrom, 2008). The details of the analysis can be found in the Methodology section. The infomap analysis for bipartite networks yields 10 consistent and statistically significant modules that we will analyse in depth in the following paragraphs. There are modules which combine many virus groups connected to very frequent gene domains, while other modules consist of one to few families connected by taxon-specific hallmark genes. The bipartite RNA plant network with the indicated modules is shown in **Figure 4.7**.

Module 1 is the most populated module in the network and consists of an assembly of +RNA viruses of several families corresponding to Branches 2 and 3 of the global RNA phylogeny, with the exception of the families *Tombusviridae, Secoviridae, Luteoviridae* and *Sobemovirus*, in Module 2. Module 1 is held together by the superfamily 1 helicase domain (S1H), which is the only gene shared by all the members of the module, and by several secondary domains, which are shared by different subgroups in the module. The capping protein methyltransferase guanydyltransferase is a common domain for many viruses; it allows viral mRNAs to be translated in the cytoplasm of eukaryotic cells (Koonin

and Moss, 2010). It is shared by viruses in different modules of the network and it is also shared by members of Module 1, except by Potyviruses —which bear the viral protein genome-linked (VPg). Filamentous viruses like *Potyviridae, Alpha/Betaflexiviridae* and *Closteroviridae* share the potyvirus-like capsid protein, and the papain-like cystein protease domain for polyprotein processing (Mann and Sanfaon, 2019; Rodamilans et al., 2018). However, contrary to the rest of potyviruses, the bipartite genus *Bymovirus* is linked to a TMV-like and TMV-readthrough capsid domains, a result that coincides with a recent publication (Kirsip and Abroi, 2019). Usually, the TMV-like capsid forms rods in benyviruses and virgaviruses (Krupovic and Koonin, 2017); two families that are also present in this module. Two superfamilies of movement proteins are grouped in Module 1: the 30K (Melcher, 2000) and the triple gene block (TGB) (Morozov and Solovyev, 2003) superfamilies. Although, not surprisingly, the analysis shows that the multipartite genera *Hordeivirus, Pomovirus* and *Goravirus* are linked to TGB domains, whereas *Furovirus* and *Tobravirus* together with the rest of monopartite genera in the *Virgaviridae* family share the 30K gene (Adams et al., 2009).

Module 2 gathers together secoviruses, ourmiaviruses, and the families *Tombusviridae* and *Luteoviridae* which are separated in branches 2 and 3 of the global RNA phylogenetic tree. They all share the single jelly roll capsid protein (SJR-CP). Module 2 is linked to Module 1 through two genes: the SJR-CP, that is shared with tymoviruses; and the chymotripsin-like protease that is shared with potyviruses. The *Secoviridae* family is the only member of the group that contains the superfamily 3 helicase domain (S3H) canonical with a picornavirus-like configuration (Thompson et al., 2014). Ourmiaviruses reveal a quimeric composition of their genomes with a SJR-CP probably of a secovirus-like origin and the MP 30K that shows resemblances to those of tombusviruses (Rastgou et al., 2009). Sobemoviruses are located in an independent module —Module 6— characterized by their own independent VPg and MPs. They are connected to module 2 through the SJR-CP and to Module 1 through the chymotrypsin-like protease, shared with potyviruses.

It is interesting how vOTU-like deubiquitinase domain (cystein protease) gathers together many modules in the network connecting +RNA, -RNA and dsRNA plant viruses. This includes several species of the order *Tymovirales* (Module 1), endornaviruses (Module 3), the genus *Tenuivirus* (Module 4), cileviruses and bluneviruses (Module 8) and the family *Aspiviridae* (Module 9).

Plant -RNA viruses of the order *Bunyavirales* (Module 4), the family *Aspiviridae* (Module 9) and the order *Mononegavirales* (rhabdoviruses in Module 7) are separated into different modules despite the high similarity found in their RdRp gene (Shi et al., 2016; Wolf et al., 2018). Viruses of the order *Bunyavirales* share taxon specific hallmark genes: the envelope protein (except for tenuiviruses, which are not enveloped), the phlebo-like nucleocapsid (Krupovic et al., 2016), the cap-snatching endonuclease —necessary for transcription (Reguera et al., 2010; Decroly et al., 2012)— and the fusion protein class II. This latter protein is integrated in the viral envelope and facilitates the fusion of membranes when the virus infects the host cell. Tenuiviruses are not enveloped and constitute a tetrapartite genus. They seem to use the fusion protein class II for a different function that is related to overcoming the insect midgut barriers (Lu et al., 2019) —in which might be an interesting example of exaptation.

Interestingly, every plant -RNA species is linked to Module 1 by the 30K MP (Mushegian and Elena, 2015). This result is consistent with the hypothesis that plant -RNA viruses were acquired from animals through HGT (Dolja and Koonin, 2011), and arthropod vectors are likely the HGT vehicles of the 30K MP that prompted the host shift from arthropods to plants (Xiong et al., 2008; Ammar et al., 2009; Whitfield et al., 2018). Additionally,

tenuiviruses are the only member of the group bearing vOTU-like deubiquitinase domain which also connects to Module 1.

Phytoreoviruses are held together into Module 5 in a very nested way including specific genes of the family *Reoviridae*. They are linked to Module 1 through the S1H domain and the capping protein methyltransferase guanydyltransferase, and to Module 2 by the S3H. Recently discovered totivirus species infecting plants (Chen et al., 2016) form an independent module despite the high similarities found with reovirus infecting fungi (Luque).

Although plant dsRNA viruses are located in different modules: Modules 3 (*Endornaviridae*), 5 (*Reoviridae*) and 10 (*Totiviridae*) share the p7-reovirus dsRNA binding domain (Masliah et al., 2013; Zhong et al., 2005).

As it was expected, the gene sharing network of plant viruses also follows a hierarchical modular organization. The modularity analysis presented here yields modules that are consistent with the taxonomical organization of the plant virus groups. In addition, some of the evolutionary relationships that are found in the reconstruction of RdRp phylogeny (Wolf et al., 2018) also emerged in this analysis as modules in the bipartite plant network. However, it is remarkable that several independent groups —attending to the RdRp phylogeny— coalesce in network modules mostly by sharing the capsid and/or helicase domains. For example, the families belonging to the "alphavirus-supergroup" are grouped in the same module as potyviruses, which appear as a separated group attending to the RdRp. Similarly, sobemoviruses form an independent module despite the similarities with potyviruses, probably as a result of the divergence of the families in genes beyond RdRp.

Overall, the analysis of the bipartite network separates families of plant viruses in a way similar to conventional phylogeny, though it integrates all-genome information and shows a principle of modularity. In particular, structural and non-strutural domains have been often shuffled across the viral groups. For example, it is shown that the capsids and helicases were horizontally transferred from group to group several times. A special case is the MP, an essential gene for plant infection. Broad horizontal spread of MPs led to major shifts in the lifestyle of viruses, especially of -RNA viruses of arthropods. Ecological interactions permitted the infection of plants after the acquisition of MPs by arthropod -RNA viruses that were relegated to become a vector in the transmission. Although further analysis must be carried out, it is likely that multipartitism favours domain shuffling. Multipartite viruses share, in general, more genes than other genome configurations. A paradigm of modularity are ourmiaviruses, whose genomes are a chimera of unrelated modules. The network analysis shows that, within a family, multipartite genera might change the capsid or movement protein, but conserve the replicative domain. In addition, multipartite viruses can acquire domains additional to those that are family-specific, as it happens with tenuiviruses, the only member of its group with vOTU-like deubiquitinase domain.

From an evolutionary perspective, the genome composition of viruses in the network is highly conditioned by the family to which the virus belongs. Consequently, the results obtained with the analysis of the RNA bipartite network are congruent with classical taxonomical groups. In this way, the adaptation to novel niches can be achieved with the adaptation of pre-existing family-specific domains to novel functions —exaptation. However, the extensive HGT across family groups suggests that extant viral families are the result of a bottom-up constructive process that may have started with an initially small group of genes and gene combinations, to be followed by subsequent expansion.

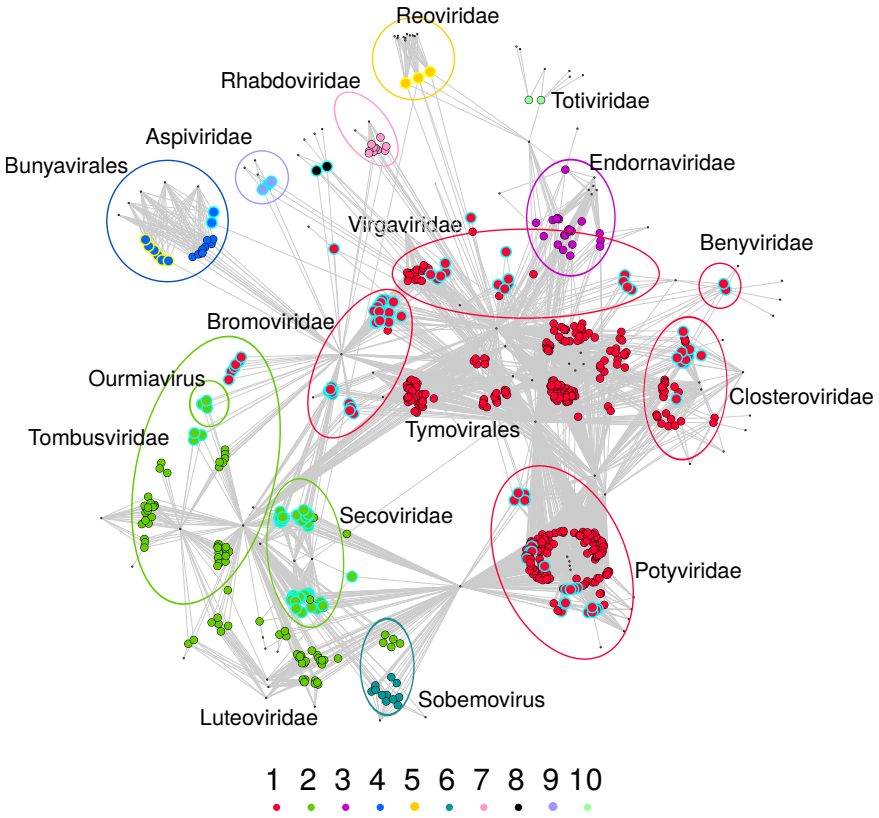

Figure 4.7: **Bipartite RNA network of plant viruses.**
Coloured nodes are genomes which are connected to genes depicted as grey dots. Different node colours are modules in the network calculated with the infomap algorithm (Rosvall and Bergstrom, 2008). The contour around the nodes describes multipartite (light blue) and segmented (yellow) genomes. Module 1 (red):*Tymovirales (Alphaflexiviridae, Betaflexiviridae, Tymoviridae), Benyviridae, Bromoviridae, Closteroviridae, Potyviridae, Tombusviridae (Umbravirus), Idaeovirus, Virgaviridae*. Module 2 (green):*Luteoviridae, Secoviridae, Tombusviridae, Ourmiavirus*. Module 3 (purple): *Endornaviridae*. Module 4 (dark blue):*Bunyavirales (Tospoviridae, Fimoviridae, Phenuiviridae (Tenuivirus))*. Module 5 (yellow): *Reoviridae*. Module 6 (turquoise): *Sobemovirus*. Module 7 (pink): *Rhabdoviridae, (Nucleorhabdovirus, Cytorhabdovirus)*. Module 8 (black): *Cilevirus* and *Blunevirus*. Module 9 (lavender): *Aspiviridae (Ophiovirus)*. Module 10 (light green): *Totiviridae*.

CHAPTER 5

DISCUSSION & PERSPECTIVES

The ultimate expansion of our knowledge of viral diversity, fostered by the last advances on virus transcriptomics (Shi et al., 2016), has just begun. Studies using those techniques have led to the discovery of new viral species at an unprecedented rate. Major gaps in virus evolution understanding will in all likelihood start to close in forthcoming years. The catalogue of multipartite species, mostly infecting plants, will benefit from the new sampling and sequencing techniques. It is foreseeable that their, in principle narrow, host-range will broaden. Actually, we already have evidence indicating that multipartitism is not limited to plants: possible multipartite candidates are infecting fungi and animals (Ghabrial et al., 2008; Olveira et al., 2009; Ladner et al., 2016).

Though general molecular advantages of multipartitism are, as of yet, a puzzle, these viruses do depend on the metabolism and the structure of plant tissues (Miyashita and Kishino, 2010; Sicard et al., 2019). However, it would be important to study the lifestyles of non-plant multipartite viruses at a level of detail similar to that of multipartite plant viruses to find out whether the complementation among genomic segments follows a mechanism analogous to that known to take place in plant tissues. Spatial structure might be essential to guarantee the persistence of multipartite viruses by reducing the cost of coinfection through local clustering. Recent theoretical work illustrates the importance of spatial propagation for the maintenance of multipartite viruses, not only in regard to their within-host transmission: actually, most multipartite species described infect monocultures where plants are spatially distributed. In particular, a structured host-to-host transmission, together with genetic drift, as been put forward as a plausible explanation for a contingent success of multipartite viruses over monopartite counterparts (Valdano et al., 2019). An

independent approach considers a spatial distribution of hosts that can be progressively infected. Intermediate infected hosts with one to few genomic segments that remain for a long enough time in the host might boost a more pervasive infection compared to the monopartite case (Zhang et al., 2019). The previous theoretical results notwithstanding, this intermediate, latent state, is not likely to exist in nature.

Different ways in which the complementation requirement can be alleviated are possible, but not yet investigated. Perhaps the success of multipartitism is partly relying on a dynamical strategy where its advantages are the opportunistic colonization of new hosts by eliciting fast adaptive responses through complementation of formerly independent genomic segments. Cooperation of such segments is a must for this strategy to be plausible. In fact, it is plants that offer a particularly suitable environment for loose cooperative associations between virus and virus-like agents of different origins. The association of genetic elements might have turned permanent when the ecological conditions drove the partners to an interdependent relationship. For example, if infection becomes conditional on the joint cooperation of a pair of genetic elements, a bipartite species appears as a natural solution. We hypothesize that this scenario offers a plausible origin for multipartite species. However, further mathematical models and additional empirical evidence will be essential to reveal which ecological conditions might favour the emergence of multipartite forms in competition with monopartite viral species.

Cooperation of genetic elements could be followed by recombination as a fundamental mechanism for the generation of novel viruses (Sachs and Bull, 2005). However, multipartite and segmented viruses apparently break that rule (Varsani et al., 2018). In particular cases, such as -RNA viruses, the coverage of the genetic material by a nucleocapsid protein represents a physical limitation to recombination, since the genetic material is not exposed. Nonetheless, the lack of recombination of the rest of Baltimore classes might be compensated by a more plastic mechanism of reassortment or gene shuffling. In fact, multipartite RNA viruses are more prone to share genes with other multipartite species than segmented or monopartite viruses with akin genome configurations. Cooperation might show up as a modular constructive principle of multipartite viral species. In fact, cooperation acquires a broader meaning if ensembles of viral species are portrayed as complex, gene-sharing networks, where it emerges as a distributed property of the ensemble. Contrary to highly stable associations of genes in chromosomes, gene sharing in multipartite viruses offers an exploding number of combinatorial possibilities that might translate into many possible different viral species in short evolutionary time.

Whether associations among genetic segments is a stable strategy or whether they might evolve to monoparticle viral forms (segmented or monopartite) remains as a further open question (Nee and Maynard-Smith, 1990). Actually, there might be restrictions to recombination and single particle encapsidation that either stabilize the multipartite state or at least delay the emergence of a monopartite cognate form —provided it would be a fitter solution. Multipartite viruses have emerged many times in evolution, and at present it seems easier to understand them as a fit, but only transiently stable solution in evolutionary time. Multipartite viruses might be the dynamic product of a huge and plastic pangenome that is constantly proposing, permitting and sustaining new associations in a complex, changing and diverse global ecology.

PART IV

CONCLUSIONES–CONCLUSIONS

CONCLUSIONES

El trabajo desarrollado en esta tesis ha dado lugar a las siguientes conclusiones principales:

1. Una búsqueda exhaustiva en bases públicas de datos genómicos muestra que la prevalencia de los virus multipartitos en la Virosfera es de alrededor del 14%. Esta abundancia apoya el hecho de que los virus multipartitos son una estrategia evolutiva estable.

2. Sobre el 90% de las especies multipartitas infectan plantas. Además, sorprendentemente, cerca de la mitad de los virus conocidos que infectan plantas son multipartitos. Las plantas ofrecen condiciones particularmente adecuadas para esta alta prevalencia de virus multipartitos.

3. Las poblaciones virales que sufren de deleciones genéticas pueden experimentar una transición hacia la multipartición genómica. Se ha demostrado que la complementación entre genomas defectivos sólo es posible a densidades virales suficientemente altas. La presión por complementarse se reduce si los virus se propagan es hospedadores espacialmente estructurados, tal y como se demuestra en el caso de los tejidos vegetales.

4. La asociación entre virus y satélites da lugar a fenotipos emergentes. Estas nuevas propiedades etiológicas y epidemiológicas pueden representar una ventaja evolutiva que contrarreste el coste de la coinfección. La asociación entre virus y satélites puede ser el primer paso hacia el surgimiento de especies multipartitas.

5. Los virus multipartitos de ARN tienen un origen evolutivo polifilético, ya que han surgido varias veces a lo largo de la filogenia de los virus de ARN.

6. El intercambio de genes entre virus multipartitos de ARN es más probable que dentro de otras configuraciones genómicas. El intercambio genético puede actuar como un principio constructivo modular para especies multipartitas y conferir una ventaja adaptativa ante cambios ambientales.

CONCLUSIONS

The work developed in this thesis leads to the following main conclusions:

1. The exhaustive search of publicly available genomic databases shows that the prevalence of multipartite viruses in the Virosphere is around 14%. Their abundance strongly supports that multipartite viruses can be considered a stable evolutionary strategy.

2. Up to 90% of multipartite species infect plants. Remarkably, about one half of all known plant viruses are multipartite. Plants must offer particularly suitable conditions for this high prevalence of multipartite viruses.

3. Viral populations that suffer from genetic deletions can experience a transition towards multipartitism. However, it had been shown that complementation of defective genomes is only possible at sufficiently high viral densities. The critical viral density can be lowered down if viruses propagate in spatially structured hosts, as it occurs within-host propagation of plant infections.

4. The association between viruses and satellites lead to emerging phenotypes. These new aetiological and epidemiological properties can represent an evolutionary advantage counteracting the cost of coinfection. The association between independent viruses and satellites can act as a stepping stone towards the rise of multipartite species.

5. Multipartite RNA viruses have a polyphyletic origin, as they have emerged many times along RNA phylogeny.

6. The sharing of genes among multipartite RNA viruses is more likely than within groups with monopartite or segmented genomes. Gene shuffling can act as a modular constructive principle for multipartite species and confer an adaptive advantage under environmental changes.

References

Adams, M. J., Antoniw, J. F., and Kreuze, J. Virgaviridae: a new family of rod-shaped plant viruses. *Arch. Virol.*, 154(12):1967–1972, 2009.

Aguirre, J. and Manrubia, S. C. Effects of spatial competition on the diversity of a quasispecies. *Phys. Rev. Lett.*, 100:38106, 2008.

Ahlquist, P. Parallels among positive-strand rna viruses, reverse-transcribing viruses and double-stranded RNA viruses. *Nat. Rev. Microbiol.*, 4:371–382, 2006.

Ahola, T. New phylogenetic grouping of positive-sense rna viruses is concordant with replication complex morphology. *mBio*, 10(4), 2019.

Ammar, E.-D., Tsai, C.-W., Whitfield, A. E., Redinbaugh, M. G., and Hogenhout, S. A. Cellular and molecular aspects of rhabdovirus interactions with insect and plant hosts. *Annual Review of Entomology*, 54(1):447–468, 2009.

Auguste, A. J., Kaelber, J. T., Fokam, E. B., Guzman, H., Carrington, C. V. F., Erasmus, J. H., Kamgang, B., Popov, V. L., Jakana, J., Liu, X., Wood, T. G., Widen, S. G., Vasilakis, N., Tesh, R. B., Chiu, W., and Weaver, S. C. A newly isolated reovirus has the simplest genomic and structural organization of any reovirus. *J. Virol.*, 89(1):676–687, 2015.

Bald, J. G. and Briggs, G. E. Aggregation pf virus particles. *Nature*, 140(111):111, 1937.

Baltimore, D. Expression of animal virus genomes. *Bacteriol Rev*, 35(3), 1971.

Bangham, C. R. and Kirkwood, T. B. Defective interfering particles and virus evolution. *Trends Microbiol*, 1(7), 1993.

Beniac, D. R., Melito, P. L., deVarennes, S. L., Hiebert, S. L., Rabb, M. J., Lamboo, L. L., Jones, S. M., and Booth, T. F. The organisation of ebola virus reveals a capacity for extensive, modular polyploidy. *PLOS ONE*, 7(1):1–12, 01 2012.

Bergua, M., Zwart, M. P., El-Mohtar, C., Shilts, T., Elena, S. F., and Folimonova, S. Y. A viral protein mediates superinfection exclusion at the whole-organism level but is not required for exclusion at the cellular level. *J Virol*, 88(19):11327–38, 2014.

Betancourt, M., Fereres, A., Fraile, A., and García-Arenal, F. Estimation of the effective number of founders that initiate an infection after aphid transmission of a multipartite plant virus. *J. Virol.*, 82(24):12416–12421, 2008.

Betancourt, M., Fraile, A., and García-Arenal, F. Cucumber mosaic virus satellite rnas that induce similar symptoms in melon plants show large differences in fitness. *J. Gen. Virol.*, 92(8):1930–1938, 2011.

Betancourt, M., Escriu, F., Fraile, A., and García-Arenal, F. Virulence evolution of a generalist plant virus in a heterogeneous host system. *Evolutionary Applications*, 6: 875–890, 2013.

Betancourt, M., Fraile, A., Milgroom, M. G., and Garca-Arenal, F. Aphid vector population density determines the emergence of necrogenic satellite rnas in populations of cucumber mosaic virus. *J. Gen. Virol.*, 97(6):1453–1457, 2016.

Boerlijst, M. C. and van Ballegooijen, W. M. Spatial pattern switching enables cyclic evolution in spatial epidemics. *PLOS Comp. Biol.*, 6(12):1–7, 12 2010.

Callister, D. M., Winter, A. D., Page, A. P., and Maizels, R. M. Four abundant novel transcript genes from toxocara canis with unrelated coding sequences share untranslated region tracts implicated in the control of gene expression. *Molecular and Biochemical Parasitology*, 162(1):60 – 70, 2008.

Campillo-Balderas, J. A., Lazcano, A., and Becerra, A. Viral genome size distribution does not correlate with the antiquity of the host lineages. *Front. Ecol. Evol.*, 3:143, 2015.

Catalán, P., Elena, S. F., Cuesta, J. A., and Manrubia, S. Parsimonious scenario for the *de novo* emergence of viroid-like replicons. *Viruses*, 2019.

Celix, A., Rodriguez-Cerezo, E., and Garcia-Arenal, F. New satellite rnas but no di rnas are found in natural populations of tomato bushy stunt tombusvirus. *Virology*, 239:277–284, 1997.

Chao, L. Levels of selection, evolution of sex in RNA viruses and the origin of life. *J. Theor. Biol.*, 153:229–246, 1991.

Chen, S., Cao, L., Huang, Q., Qian, Y., and Zhou, X. The complete genome sequence of a novel maize-associated totivirus. *Arch. Virol.*, 161(2):487–490, 2016.

Conant, G. C. and Wolfe, K. H. Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.*, 9:938–950, 2008.

Cotmore, S. F., Agbandje-McKenna, M., Chiorini, J. A., Mukha, D. V., Pinteland, D. J., Qiu, J., Soderlund-Venermo, M., Tattersall, P., Tijssen, P., Gatherer, D., and Davison, A. J. The family *parvoviridae*. *Arch Virol*, 159(5), 2014.

Cotton, S., Grangeon, R., Thivierge, K., Mathieu, I., Ide, C., Wei, T., Wang, A., and Laliberté, J.-F. Turnip mosaic virus RNA replication complex vesicles are mobile, align with microfilaments, and are each derived from a single viral genome. *J Virol*, 83(20): 10460–10471, 2009.

Cros, J. F. and Palese, P. Trafficking of viral genomic rna into and out of the nucleus: influenza, thogoto and borna disease viruses. *Virus Res.*, 95(1):3 – 12, 2003.

Cuesta, J. A. and Manrubia, S. Enumerating secondary structures and structural moieties for circular RNAs. *J. Theor. Biol.*, submitted, 2016.

Cuesta, J. A., Aguirre, J., Capitán, J. A., and Manrubia, S. C. The struggle for space: Viral extinction through competition for cells. *Phys. Rev. Lett.*, 106:028104, 2011.

Dall'Ara, M., Ratti, C., Bouzoubaa, S. E., and Gilmer, D. Ins and outs of multipartite positive-strand rna plant viruses: Packaging versus systemic spread. *Viruses*, 8:228, 2016.

Dawson, W. O. and Hilf, M. E. Host–range determinants of plant viruses. *Annual Reviews of Plant Physiology Plant Molecular Biology*, 43:527–55, 1992.

Decroly, E., Ferron, Fran c., Lescar, J., and Canard, B. Conventional and unconventional mechanisms for capping viral mRNA. *Nat. Rev. Microbiol.*, 10:51–65, 2012.

Derrick, K., Brlansky, R., da Graca, J., Lee, R., Timmer, L., and Nguyen, T. Partial characterization of a virus associated with citrus ringspot. *Phytopathology*, 78:1298–1301, 1988.

Desbiez, C., Moury, B., and Lecoq, H. The hallmarks of green viruses: Do plant viruses evolve differently from the others? *Infection, Genetics and Evolution*, 11(5):812 – 824, 2011.

Deshoux, M., Monsion, B., and Uzest, M. Insect cuticular proteins and their role in transmission of phytoviruses. *Curr. Op. Virol.*, 33:137 – 143, 2018.

Dietzgen, R. G., Mann, K. S., and Johnson, K. N. Plant virus-insect vector interactions: current and potential future research directions. *Viruses*, 8:303, 2016.

Dodds, J. A. Satellite tobacco mosaic virus. *Ann. Rev. Phytopathol.*, 36:295–310, 1998.

Dolja, V. V. and Koonin, E. V. Common origins and host-dependent diversity of plant and animal viromes. *Curr. Op. Virol.*, 184(5):322–331, 2011.

Dolja, V. V. and Koonin, E. V. Metagenomics reshapes the concepts of rna virus evolution by revealing extensive horizontal virus transfer. *Virus Research*, 244:36 – 52, 2018.

Domingo, E., Sabo, D., and Weissmann, C. Nucleotide sequence heterogeneity of an RNA phage population. *Cell*, 13:735–744, 1978.

Drake, J. W. and Holland, J. J. Mutation rates among rna viruses. *Proc Natl Acad Sci USA*, 96(24):13910–13913, 1999.

Dunham, J., Simmons, H., Holmes, E., and Stephenson, A. Analysis of viral (zucchini yellow mosaic virus) genetic diversity during systemic movement through a cucurbita pepo vine. *Virus Research*, 191:172 – 179, 2014.

Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523, 1971.

El Omari, K., Sutton, G., J Ravantti, J., Zhang, H., Walter, T., Grimes, J., H Bamford, D., Stuart, D., and Mancini, E. Plate tectonics of virus shell assembly and reorganization in phage phi-8, a distant relative of mammalian reoviruses. *Structure*, 21, 2013.

Elena, S. F., Bernet, G. P., and Carrasco, J. L. The games plant viruses play. *Current Opinion in Virology*, 8:62–67, 2014.

Falk, B. W. and Tsai, J. H. Biology and molecular biology of viruses in the genus *tenuiviruses*. *Annu. Rev. Phytopathol.*, 36:139–163, 1998.

Farci, P., Karayiannis, P., Lai, M. E., Marongiu, F., Orgiana, G., Balestrieri, A., and Thomas, H. C. Acute and chronic hepatitis delta virus infection: direct or indirect effect on hepatitis b virus replication? *J Med Virol.*, 26(3):279–288, 1988.

Fargette, D., Pinel-Galzi, A., Sérémé, D., Lacombe, S., Hébrard, E., Traoré, O., and Konaté, G. Diversification of *rice yellow mottle virus* and related viruses spans the history of agriculture from the neolithic to the present. *PLoS Pathogens*, 4:e1000125, 2008.

Ferrero, D., Ferrer-Orta, C., and N, V. *Viral RNA–Dependent RNA Polymerases: A Structural Overview*, volume 88. Springer, Singapore, 2018.

Fischer, M. . G. . Sputnik and mavirus: more than just satellite viruses. *NRM*, 10(1), 2011.

Flores, R., Grubb, D., Elleuch, A., Nohales, M.-A., Delgado, S., and Gago, S. Rolling-circle replication of viroids, viroid-like satellita rnas and hepatitis delta virus: Variations on a theme. *RNA Biology*, 8:200–206, 2011.

French, R. and Ahlquist, P. Characterization and engineering of sequences controlling in vivo synthesis of brome mosaic virus subgenomic rna. *J Virol*, 62(7):2411–2420, 1988.

Fulton, R. W. The effect of dilution on necrotic ringspot virus infectivity and the enhancement of infectivity by noninfective virus. *Virology*, 18(3):477 – 485, 1962.

Fulton, R. W. Biological significance of multicomponent viruses. *Ann. Rev. Phytopathol.*, 18:131–146, 1980.

Galasso, G. J. Quantitative studies on the quality, effects of aggregation and thermal inactivation of vesicular stomatitis virus. *Archiv für die gesamte Virusforschung*, 21(3): 437–446, 1967.

Gallet, R., Fabre, F., Thébaud, G., Sofonea, M. T., Sicard, A., Blanc, S., and Michalakis, Y. Small bottleneck size in a highly multipartite virus during a complete infection cycle. *J. Virol.*, 92(14), 2018a.

Gallet, R., Michalakis, Y., and Blanc, S. Vector-transmission of plant viruses and constraints imposed by virusvector interactions. *Curr. Op. Virol.*, 33:144 – 150, 2018b.

García-Arriaza, J., Manrubia, S. C., Toja, M., Domingo, E., and Escarmís, C. Evolutionary transition toward defective RNAs that are infectious by complementation. *Journal of Virology*, 78:11678–11685, 2004.

Geigenmüller-Gnirke, U., Weiss, B., Wright, R., and Schlesinger, S. Complementation between sindbis viral RNAs produces infectious particles with a bipartite genome. *Proc. Natl. Acad. Sci. USA*, 88:3253, 1991.

Geoghegan, J. L., Duchłne, S., and Holmes, E. C. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLOS Pathogens*, 13(2):1–17, 02 2017.

Ghabrial, S., Castón, J., Coutts, R., Hillman, B., Jiang, D., D-H., K., and Moriyama, H. ICTV Virus taxonomy profiles: Chrysoviridae. *J. Gen. Virol.*, 99:19 – 20, 2008.

Gibbs, A. J., Ohshima, K., Phillips, M. J., and Gibbs, M. J. The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. *PLoS ONE*, 3:e2523, 2008.

Gibbs, A. J., Fargette, D., Garcí-Arenal, F., and Gibbs, M. J. Time – the emerging dimension of plant virus studies. *J. Gen. Virol.*, 91:13–22, 2010.

Gibbs, A. J., Wood, J., Garcia-Arenal, F., Ohshima, K., and Armstrong, J. S. Tobamoviruses have probably co-diverged with their eudicotyledonous hosts for at least 110 million years. *Virus Evolution*, 1(1), 2015.

Gilbert, C. and Cordaux, R. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Curr Op Virol*, 25:16 – 22, 2017.

Gladyshev, E. A. and Arkhipova, I. R. A widespread class of reverse transcriptase-related cellular genes. *Proc. Natl. Acad. Sci.*, 108(51):20311–20316, 2011.

Goldbach, R. W. Molecular evolution of plant RNA viruses. *Annu. Rev. Phytopathol.*, 24: 289–310, 1986.

Gonzalez-Jara, P., Fraile, A., Canto, T., and Garca-Arenal, F. The multiplicity of infection of a plant virus varies during colonization of its eukaryotic host. *J. Virol.*, 83(15):7487–7494, 2009.

Gopinath, K. and Kao, C. C. Replication-independent long-distance trafficking by viral rnas in nicotiana benthamiana. *The Plant Cell*, 19(4):1179–1191, 2007.

Grande-Pérez, A., Lázaro, E., Domingo, E., and Manrubia, S. C. Suppression of viral infectivity through lethal defection. *Proc. Natl. Acad. Sci. USA*, 102:4448–4452, 2005.

Grigoras, I., Timchenko, T., and Gronenborn, B. Transcripts encoding the nanovirus master replication initiator proteins are terminally redundant. *J Gem Virol*, 89:583–593, 2008.

Grigoras, I., Timchenko, T., Katul, L., Grande-Pérez, A., Vetten, H.-J., and Gronenborn, B. Reconstitution of authentic nanovirus from multiple cloned dnas. *J. Virol.*, 83(20): 10778–10787, 2009.

Grigoras, I., Timchenko, T., Grande-Pérez, A., Katul, L., Vetten, H.-J., and Gronenborn, B. High variability and rapid evolution of a nanovirus. *J Virol*, 84(18):9105–9117, 2010.

Gutiérrez, S., Yvon, M., Thébaud, G., Monsion, B., Michalakis, Y., and Blanc, S. Dynamics of the multiplicity of cellular infection in a plant virus. *PLoS Pathog.*, 6:e1001113, 2010.

Gutierrez, S., Michalakis, Y., and Blanc, S. Virus population bottlenecks during within-host progression and host-to-host transmission. *Curr. Op. Virol.*, 2(5):546 – 555, 2012.

Gutierrez, S., Pirolles, E., Yvon, M., Baecker, V., Michalakis, Y., and Blanc, S. The multiplicity of cellular infection changes depending on the route of cell infection in a plant virus. *J. Virol.*, 89(18):9665–9675, 2015.

Haber, S., Ikegami, M., Bajet, N., and Goodman, R. Evidence for a divided genome in bean golden mosaic virus, a geminivirus. *Nature*, 289:324–326, 1981.

Hadid, A., Flores, R., Randles, J. W., and Palukaitis, P. *Viroids and Satellites*. Elsevier Science, 1st edition, 2017.

Hajimorad, M. R., Kurath, G., Randles, J. W., and Francki, R. I. B. Change in phenotype and encapsidated rna segments of an isolate of alfalfa mosaic virus: An influence of host passage. *Journal of General Virology*, 72(12):2885–2893, 1991.

Hajimorad, M. R., Ghabrial, S. A., and Roossinck, M. J. De novo emergence of a novel satellite rna of cucumber mosaic virus following serial passages of the virus derived from rna transcripts. *Archives of Virology*, 154(1):137–140, 2009.

Harrison, B. D., Kubo, S., Robinson, D. J., and Hutcheson, A. M. The multiplication cycle of tobacco rattle virus in tobacco mesophyll protopasts. *J. Gen. Virol.*, 33:237, 1976.

Hayakawa, T., Kojima, K., Nonaka, K., Nakagaki, M., Sahara, K., ichiro Asano, S., Iizuka, T., and Bando, H. Analysis of proteins encoded in the bipartite genome of a new type of parvo-like virus isolated from silkworm – structural protein with DNA polymerase motif. *Virus Res.*, 66:101–108, 2000.

Heinlein, M., Padgett, H. S., Gens, J. S., Pickard, B. G., Casper, S. J., Epel, B. L., and Beachy, R. N. Changing patterns of localization of the tobacco mosaic virus movement protein and replicase to the endoplasmic reticulum and microtubules during infection. *The Plant Cell*, 10(7):1107–1120, 1998.

Holmes, E. C. The expanding virosphere. *Cell Host & microbe*, 20:279–280, 2016.

Hu, C.-C., Hsu, Y.-H., YH, H., Lin, N.-S., and NS, L. Satellite rnas and satellite viruses of plants. *Viruses*, 1(3), 2009.

Hu, Z., Zhang, X., Liu, W., Zhou, Q., Zhang, Q., Li, G., and Yao, Q. Genome segments accumulate with different frequencies in *bombyx mori bidensovirus*. *J. Basic Microbiol.*, 56:1338–1343, 2016.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., and Le Mercier, P. Viralzone: a knowledge resource to understand virus diversity. 39(suppl 1):D576–D582, 2011.

Iltis, R., Bald, J., Schneider, S., and Dawson, W. Logistic and poisson models for infection by multicomponent plant viruses. *J of Virol Methods*, 26(2):147 – 157, 1989.

Iranzo, J. and Manrubia, S. C. Evolutionary dynamics of genome segmentation in multipartite viruses. *Proc. Royal Soc. London B*, 279(1743):3812–3819, 2012. doi: 10.1098/rspb.2012.1086.

Iranzo, J., Koonin, E. V., Prangishvili, D., and Krupovic, M. Bipartite network analysis of the archaeal virosphere: Evolutionary connections between viruses and capsidless mobile elements. *J. Virol.*, 90:11043–11055, 2016a.

Iranzo, J., Krupovic, M., and Koonin, E. V. The double-stranded dna virosphere as a modular hierarchical network of gene sharing. *mBio*, 7:e00978, 2016b.

Iranzo, J., Koonin, E. V., and Krupovic, M. A network perspective on the virus world. *Communicative and integrative biology*, 10(2):e1296614, 2017.

Kaido, M., Tsuno, Y., Mise, K., and Okuno, T. Viral cell-to-cell movement requires formation of cortical punctate structures containing Red clover necrotic mosaic virus movement protein. *Virology*, 413:205–15, 2011.

Kassanis, B. Properties and behaviour of a virus depending for its multiplication on another. *J. Gen. Microbiol.*, 27:477–488, 1962.

Kawakami, S., Watanabe, Y., and Beachy, R. N. Tobacco mosaic virus infection spreads cell to cell as intact replication complexes. *Proc. Natl. Acad. Sci. USA*, 101(16):6291–6296, 2004.

Kazlauskas, D., Dayaram, A., Kraberger, S., Goldstien, S., Arvind, and Krupovic, M. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology*, 504:114 – 121, 2017.

Keeling, P. J. and Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.*, 9:605–618, 2008.

Khoshnan, A. and Alderete, J. F. Characterization of double-stranded rna satellites associated with the *trichomonas vaginalis* virus. *J Virol*, 69(11), 1995.

Kim, K. H., Narayanan, K., and Makino, S. Assembled coronavirus from complementation of two defective interfering RNAs. *J. Virol.*, 71:3922, 1997.

Kim, Y. J., Kim, J. Y., Kim, J. H., Yoon, S. M., Yoo, Y.-B., and Yie, S. W. The identification of a novel pleurotus ostreatus dsrna virus and determination of the distribution of viruses in mushroom spores. *The Journal of Microbiology*, 46(1):95–99, 2008.

King, A. M., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J., editors. *The Subviral Agents*. Elsevier, San Diego, 2012.

Kirkwood, T. B. L. and Bangham, C. R. Cycles, chaos, and evolution in virus cultures: a model of defective interfering particles. *Proc Natl Acad Sci USA*, 91:8685–89, 1994.

Kirsip, H. and Abroi, A. Protein structure-guided hidden markov models (HMMs) as a powerful method in the detection of ancestral endogenous viral elements. *Viruses*, 11 (4):320, 2019.

Knipe, D. M. and Howley, P. M., editors. *Fields Virology*, volume I. William Lippincott & Wilkins, Philadelphia, USA, 5th edition, 2007.

Koonin, E., Gorbalenya, A., and Chumakov, K. Tentative identification of rna-dependent rna polymerases of dsrna viruses and their relationship to positive strand rna viral polymerases. *FEBS Letters*, 252(1):42 – 46, 1989.

Koonin, E. V. The turbulent network dynamics of microbial evolution and the statistical tree of life. *J. Mol. Evol.*, 80:244–250, 2015.

Koonin, E. V. Viruses and mobile elements as drivers of evolutionary transitions. *Philos Trans R Soc Lond B Biol Sci*, 371(1701):20150442, 2016.

Koonin, E. V. and Dolja, V. Evolution and taxonomy of positive–strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit. Rev. Biochem. Mol. Biol.*, 28:375–430, 1993.

Koonin, E. V. and Dolja, V. V. A virocentric perspective on the evolution of life. *Curr Op Virol*, 3(5):546 – 557, 2013.

Koonin, E. V. and Dolja, V. V. Virus world as an evolutionary network of viruses and capsidless selfish elements. 78(2):278–303, 2014.

Koonin, E. V. and Moss, B. Viruses know more than one way to don a cap. *Proc. Natl. Acad. Sci. USA*, 107(8):3283–3284, 2010.

Koonin, E. V., Senkevich, T. G., and Dolja, V. The ancient virus world and evolution of cells. *Biol. Direct*, 1:29, 2006.

Koonin, E. V., Wolf, Y. I., Nagasaki, K., and Dolja, V. V. The big bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat. Rev. Microbiol.*, 6:925–939, 2008.

Koonin, E. V., Dolja, V. V., and Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*, 479480:2–25, 2015.

Kormelink, R., Garcia, M. L., Goodin, M., Sasaya, T., and Haenni, A.-L. Negative-strand RNA viruses: The plant-infecting counterparts. *Virus Research*, 162(12):184 – 202, 2011.

Krichevsky, A., Kozolovsky, S. V., Gafni, Y., and Citovsky, V.

Krishnamurthy, S. R. and Wang, D. Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. *Virology*, 516:108 – 114, 2018.

Krupovic, M. and Cvirkaite-Krupovic, V. Virophages or satellite viruses? *NRM*, 9(11): 762 – 3, 2011.

Krupovic, M. and Cvirkaite-Krupovic, V. Towards a more comprehensive classification of satellite viruses. *NRM*, 234(10), 2012.

Krupovic, M. and Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci. USA*, 114(12):E2401–E2410, 2017.

Krupovic, M., Kuhn, J. H., and Fisher, M. G. . A classification system for virophages and satellite virus. *Arch. Virol.*, 161(1):233–247, 2016.

Kuchibhatla, D. B., Sherman, W. A., Chung, B. Y. W., Cook, S., Schneider, G., Eisenhaber, B., and Karlin, D. G. Powerful sequence similarity search methods and in depth manual analyses can identify remote homologs in many apparently orphan viral proteins. *J. Virol.*, 88:10–20, 2014.

Ladner, J. T., Wiley, M. R., Beitzel, B., Auguste, A. J., II, A. P. D., Lindquist, M. E., Sibley, S. D., Kota, K. P., Fetterer, D., Eastwood, G., Kimmel, D., Prieto, K., Guzman, H., Aliota, M. T., Reyes, D., Brueggemann, E. E., John, L. S., Hyeroba, D., Lauck, M., Friedrich, T. C., O'Connor, D. H., Gestole, M. C., Cazares, L. H., Popov, V. L., Castro-Llanos, F., Kochel, T. J., Kenny, T., White, B., Ward, M. D., Loaiza, J. R., Goldberg, T. L., Weaver, S. C., Kramer, L. D., Tesh, R. B., and Palacios, G. A multicomponent animal virus isolated from mosquitoes. *Cell Host & Microbe*, 20:357, 2016.

Lago, M., Rodríguez, J. F., Bandín, I., and Dopazo, C. P. Aquabirnavirus polyploidy: a new strategy to modulate virulence? *J. Gen. Virol.*, 97(5):1168–1177, 2016.

Laliberté, J.-F. and Sanfaçon, H. Cellular remodeling during plant virus infection. *Annu. Rev. Phythopathol.*, 48(1):69–91, 2010.

Laliberté, J.-F. and Zheng, H. Viral manipulation of plant host membranes. *Annu. Rev. Virol.*, 1(1):237–259, 2014.

Lamprecht, R. L., Spaltman, M., Stephan, D., Wetzel, T., and Burger, J. T. Complete nucleotide sequence of a south african isolate of grapevine fanleaf virus and its associated satellite rna. *Viruses*, 5(7):1815–1823, 2013.

Lane, L. *The RNAs of multipartite and satellite viruses of plants*, volume 2. CRC Press, 1979.

Lefeuvre, P. and Moriones, E. Recombination as a motor of host switches and virus emergence: geminiviruses as case studies. *Curr Op Virol*, 10:14 – 19, 2015.

Lefeuvre, P., Martin, D. P., Elena, S. F., Shepherd, D. N., and Roumagnac, P. Evolution and ecology of plant viruses. *Nat. Rev. Microbiol.*, 17:632–644, 2019.

Lefkowitz, E. J., Adams, M. J., Davison, A. J., Siddell, S. G., and Simmonds, P., editors. *Virus Taxonomy: The Classification and Nomenclature of Viruses. The online 10th Report of the ICTV*. EC 47, London, UK, 2015.

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E. C., and Zhang, Y.-Z. Unprecedented genomic diversity of rna viruses in arthropods reveals the ancestry of negative-sense rna viruses. *eLife*, 4:e05378, 2015.

Lister, R. M. Possible relationships of virus-specific products of tobacco rattle virus infections. *Virology*, 28:350–353, 1966.

Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., Peng, Y., Ghabrial, S. A., and Yi, X. Widespread horizontal gene transfer from double-stranded rna viruses to eukaryotic nuclear genomes. *J Virol*, 84(22):11876–11887, 2010.

Lu, G., Li, S., Zhou, C., Qian, X., Xiang, Q., Yang, T., Wu, J., Zhou, X., Zhou, Y., Ding, X. S., and Tao, X. Tenuivirus utilizes its glycoprotein as a helper component to overcome insect midgut barriers for its circulative and propagative transmission. *PLOS Pathogens*, 15(3):1–29, 03 2019.

Luque, D. a.

Madinda, N. F., Ehlers, B., Wertheim, J. O., Akoua-Koffi, C., Bergl, R. A., Boesch, C., Akonkwa, D. B. M., Eckardt, W., Fruth, B., Gillespie, T. R., Gray, M., Hohmann, G., Karhemere, S., Kujirakwinja, D., Langergraber, K., Muyembe, J.-J., Nishuli, R., Pauly, M., Petrzelkova, K. J., Robbins, M. M., Todd, A., Schubert, G., Stoinski, T. S., Wittig, R. M., Zuberbühler, K., Peeters, M., Leendertz, F. H., and Calvignac-Spencer, S. Assessing host-virus codivergence for close relatives of merkel cell polyomavirus infecting african great apes. *J. Virol.*, 90(19):8531–8541, 2016.

Maes, P., Alkhovsky, S. V., Bào, Y., Beer, M., Birkhead, M., Briese, T., Buchmeier, M. J., Calisher, C. H., Charrel, R. N., Choi, I. R., Clegg, C. S., de la Torre, J. C., Delwart, E., DeRisi, J. L., Di Bello, P. L., Di Serio, F., Digiaro, M., Dolja, V. V., Drosten, C., Druciarek, T. Z., Du, J., Ebihara, H., Elbeaino, T., Gergerich, R. C., Gillis, A. N., Gonzalez, J.-P. J., Haenni, A.-L., Hepojoki, J., Hetzel, U., H, T., Hóng, N., Jain, R. K., Jansen van Vuren, P., Jin, Q., Jonson, M. G., Junglen, S., Keller, K. E., Kemp, A., Kipar, A., Kondov, N. O., Koonin, E. V., Kormelink, R., Korzyukov, Y., Krupovic, M., Lambert, A. J., Laney, A. G., LeBreton, M., Lukashevich, I. S., Marklewitz, M., Markotter, W., Martelli, G. P., Martin, R. R., Mielke-Ehret, N., Mühlbach, H.-P., Navarro, B., Ng, T. F. F., Nunes, M. R. T., Palacios, G., Paweska, J. T., Peters, C. J., Plyusnin, A., Radoshitzky, S. R., Romanowski, V., Salmenperä, P., Salvato, M. S., Sanfaçon, H., Sasaya, T., Schmaljohn, C., Schneider, B. S., Shirako, Y., Siddell, S., Sironen, T. A., Stenglein, M. D., Storm, N., Sudini, H., Tesh, R. B., Tzanetakis, I. E., Uppala, M., Vapalahti, O., Vasilakis, N., Walker, P. J., Wáng, G., Wáng, L., Wáng, Y., Wèi, T., Wiley, M. R., Wolf, Y. I., Wolfe, N. D., Wú, Z., Xú, W., Yang, L., Yāng, Z., Yeh, S.-D., Zhāng, Y.-Z., Zhèng, Y., Zhou, X., Zhū, C., Zirkel, F., and Kuhn, J. H. Taxonomy of the family arenaviridae and the order bunyavirales: update 2018. *Archives of Virology*, 163(8):2295–2310, 2018.

Makino, S., Chang, M.-F., Shieh, C.-K., Kamahora, T., Vannier, D. M., Govindarajan, S., and Lai, M. M. C. Molecular cloning and sequencing of a human hepatitis delta (δ) virus rna. *Nature*, 329:343–346, 1987.

Mandahar, C. *Multiplication of RNA plant viruses*. Springer, Netherlands, 2006.

Mann, K. S. and Sanfaon, H. Expanding repertoire of plant positive-strand rna virus proteases. *Viruses*, 11(1):66, 2019.

Manrubia, S. C. and Lázaro, E. Viral evolution. *Phys. Life Rev.*, 3:65–92, 2006.

Manrubia, S. C., Domingo, E., and Lázaro, E. Pathways to extinction – beyond the error threshold. *Phil. Trans. R. Soc.*, 365:1943–1952, 2010.

Mansoor, S., Briddon, R. W., Zafar, Y., and Stanley, J. Geminivirus disease complexes: an emerging threat. *Trends in Plant Science*, 8(3):128 – 134, 2003.

Maruyama, S. R., Castro-Jorge, L. A., Ribeiro, J. M. C., Gardinassi, L. G., Garcia, G. R., Brandao, L. G., Rodrigues, A. R., Okada, M. I., Abrao, E. P., Ferreira, B. R., da Fonseca, B. A. L., , and de Miranda-Santos, I. K. F. Characterisation of divergent flavivirus ns3 and ns5 protein sequences detected in rhipicephalus microplus ticks from brazil. *Mem Inst Oswaldo Cruz*, 109(1):38–50, 2014.

Masliah, G., Barraud, P., and Allain, F. H. T. Rna recognition by double-stranded rna binding domains: a matter of shape and sequence. *Cellular and Molecular Life Sciences*, 70(11):1875–1895, 2013.

Maynard Smith, J. *On Evolution*. Edinburgh University Press, 1972.

McDonald, S. S., Nelson, M. I., Turner, P. E., and Patton, J. T. Reassortment in segmented rna viruses: mechanisms and outcomes. *Nat. Rev. Microbiol.*, 14:448–460, 2016.

Melcher, U. The 30k superfamily of viral movement proteins. *J. Gen. Virol.*, 81(1):257–266, 2000.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. Linking virus genomes with host taxonomy. *Viruses*, 8(3), 2016.

Miyashita, S. and Kishino, H. Estimation of the size of genetic bottlenecks in cell-to-cell movement of soil-borne wheat mosaic virus and the possible role of the bottlenecks in speeding up selection of variations in trans-acting genes or elements. *J Virol*, 84(4): 1828–1837, 2010.

Moreno, E., Gallego, I., Gregori, J., Lucía-Sanz, A., Soria, M. E., Castro, V., Beach, N. M., Manrubia, S., Quer, J., Esteban, J. I., Rice, C. M., Gómez, J., Gastaminza, P., Domingo, E., and Perales, C. Internal disequilibria and phenotypic diversification during replication of hepatitis c virus in a noncoevolving cellular environment. *J Virol*, 91(10), 2017.

Moreno, P., Ambrós, S., Albiach-Martí, M. R., Guerri, J., and Peña, L. Citrus tristeza virus: a pathogen that changed the course of the citrus industry. *Molecular Plant Pathology*, 9 (2):251–268, 2008.

Morozov, S. Y. and Solovyev, A. G. Triple gene block: modular design of a multifunctional machine for plant virus movement. *J. Gen. Virol.*, 84(6):1351–1366, 2003.

Moury, B., Fabre, F., and Senoussi, R. Estimation of the number of virus particles transmitted by an insect vector. *Proc. Natl. Acad. Sci. USA*, 104(45):17891–17896, 2007.

Murant, A. F. Dependence of groundnut rosette virus on its satellite rna as well as on groundnut rosette assistor luteovirus for transmission by *aphis craccivora*. *J Gen Virol*, 71:2163–66, 1990.

Mushegian, A. R. and Elena, S. F. Evolution of plant virus movement proteins from the 30k superfamily and of their homologs integrated in plant genomes. *Virology*, 476:304 – 315, 2015.

Nakashima, N., Kawahara, N., Omura, T., and Noda, H. Characterization of a novel satellite virus and a strain of himetobi p virus (dicistroviridae) from the brown planthopper, nilaparvata lugens. *Journal of Invertebrate Pathology*, 91(1):53 – 56, 2006.

Nault, L. R. Arthropod Transmission of Plant Viruses: a New Synthesis. *Annals of the Entomological Society of America*, 90(5):521–541, 1997.

Nee, S. Evolution of multicompartmental genomes in viruses. *Journal of Molecular Evolution*, 25:277–281, 1987.

Nee, S. Mutualism, parasitism and competition in the evolution of coviruses. *Phil. Trans. R. Soc. Lond. B*, 355:1607–1613, 2000.

Nee, S. The evolutionary ecology of molecular replicators. *R. Soc. open sci.*, 3:160235, 2016.

Nee, S. and Maynard-Smith, J. M. The evolutionary biology of molecular parasites. *Parasitology*, 100:S5–S18, 1990.

Ng, K. K. S., Arnold, J. J., and Cameron, C. E. Structure-function relationships among rna-dependent rna polymerases. *Curr Top Microbiol Immunol*, 320:137–56, 2008.

Nibert, M. L., Ghabrial, S. A., Maiss, E., Lesker, T., Vainio, E. J., Jiang, D., and Suzuki, N. Taxonomic reorganization of family *partitiviridae* and other recent progress in partitivirus research. *Virus Res.*, 188:128–141, 2014.

Niehl, A. and Heinlein, M. Cellular pathways for viral transport through plasmodesmata. *Protoplasma*, 248(1):75–99, 2010.

Nixon, H. L. As estimate of the number of tobacco mosaic virus particles in a single hair cell. 2(1):126–128.

Ojosnegros, S., Garca-Arriaza, J., Escarmís, C., Manrubia, S. C., Perales, C., Arias, A., Mateu, M. G., and Domingo, E. Viral genome segmentation can result from a trade-off between genetic content and particle stability. *PLoS Genet.*, 7:e1001344, 2011.

Olveira, J. G., Souto, S., Dopazo, C. P., Thiéry, R., Barja, J. L., and Bandín, I. Comparative analysis of both genomic segments of betanodaviruses isolated from epizootic outbreaks in farmed fish species provides evidence for genetic reassortment. *J. Gen. Virol.*, 90: 2940–2951, 2009.

O'Neill, F. J., Maryon, E. B., and Carroll, D. Isolation and characterization of defective simian virus 40 genomes which complement for infectivity. *J. Virol.*, 43:18–25, 1982.

Ortín, J. and Martín-Benito, J. The RNA synthesis machinery of negative-stranded RNA viruses. *Virology*, 479–480:532 – 544, 2015.

Pagan, I. and Holmes, E. C. Long-term evolution of the luteoviridae: time scale and mode of virus speciation. *J. Virol.*, 84:6177–6187, 2010.

Pantaleo, V. and Burgyán, J. Cymbidium ringspot virus harnesses rna silencing to control the accumulation of virus parasite satellite rna. *J Virol*, 82(23):11851–11858, 2008.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.-M., and Abergel, C. Pandoraviruses: Amoeba viruses with genomes up to 2.5 mb reaching that of parasitic eukaryotes. *Science*, 341(6143):281–286, 2013.

Power, A. G. Insect transmission of plant viruses: a constraint on virus variability. *Curr. Op. Plant Biol.*, 3(4):336 – 340, 2000.

Power, A. G. *Community Ecology of Plant Viruses*, pages 15–26. Springer Verlag, 2008.

Power, A. G. and Flecker, A. S. *The role of vector diversity in disease dynamics*, pages 30–47. Princeton University Press, 2010.

Pressing, J. and Reanney, D. C. Divided genomes and intrinsic noise. *J. Mol. Evol.*, 20: 135–146, 1984.

Price, W. and Spencer, E. Accuracy of the local lesion method for measuring virus activity. ii. tobacco necrosis, alfalfa mosaic, and tobacco ringspot viruses. *Am J Bot*, 30:340–346, 1943.

Pruss, G., Ge, X., Shi, X. M., Carrington, J. C., and Bowman Vance, V. Plant viral synergism: the potyviral genome encodes a broad-range pathogenicity enhancer that transactivates replication of heterologous viruses. *The Plant Cell*, 9(6):859–868, 1997.

Qin, X.-C., Shi, M., Tian, J.-H., Lin, X.-D., Gao, D.-Y., He, J.-R., Wang, J.-B., Li, C.-X., Kang, Y.-J., Yu, B., Zhou, D.-J., Xu, J., Plyusnin, A., Holmes, E. C., and Zhang, Y.-Z. A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors. *Proc. Natl. Acad. Sci. USA*, 111:6744–6749, 2014.

Rager, M., Vongpunsawad, S., Duprex, W. P., and Cattaneo, R. Polyploid measles virus with hexameric genome length. *The EMBO Journal*, 21(10):2364–2372, 2002.

Rahim, M. D., Andika, I. B., Han, C., Kondo, H., and Tamada, T. RNA4-encoded p31 of beet necrotic yellow vein virus is involved in efficient vector transmission, symptom severity and silencing suppression in roots. *J. Gen. Virol.*, 88(5):1611–1619, 2007.

Ramrez, B.-C. and Haenni, A.-L. Molecular biology of tenuiviruses, a remarkable group of plant viruses. *J Gen Virol*, 75(3):467–475, 1994.

Rasheed, M. S., Selth, L. A., Koltunow, A. M., Randles, J. W., and Rezaian, M. A. Single-stranded dna of tomato leaf curl virus accumulates in the cytoplasm of phloem cells. *Virology*, 348(1):120 – 132, 2006.

Rastgou, M., Habibi, M. K., Izadpanah, K., Masenga, V., Milne, R. G., Wolf, Y. I., Koonin, E. V., and Turina, M. Molecular characterization of the plant virus genus Ourmiavirus and evidence of inter-kingdom reassortment of viral genome segments as its possible route of origin. *J. Gen. Virol.*, 90(10):2525–2535, 2009.

Reanney, D. C. The evolution of RNA viruses. *Ann. Rev. Microbiol.*, 36:47–73, 1982.

Reguera, J., Weber, F., and Cusack, S. Bunyaviridae rna polymerases (l-protein) have an n-terminal, influenza-like endonuclease domain, essential for viral cap-dependent transcription. *PLOS Pathogens*, 6(9):1–14, 09 2010.

Reguera, J., Gerlach, P., and Cusack, S. Towards a structural understanding of RNA synthesis by negative strand RNA viral polymerases. *Curr. Op. Struct. Biol.*, 36:75 – 84, 2016.

Reiko Kikuno, Hidoyuki Toh, H. H. T. M. Sequence similarity between putative gene products of geminiviral DNAs. *Nature*, 308:562, 1984.

Ribière, M., Olivier, V., and Blanchard, P. Chronic bee paralysis: A disease and a virus like no other? *J Invert Pathol*, 103:S120 – S131, 2010.

Richards, K. E. and Tamada, T. Mapping functions on the multipartite genome of beet necrotic yellow vein virus. *Annu. Rev. Phytopathol.*, 30(1):291–313, 1992.

Roberts, S. and Stanley, J. Lethal mutations within the conserved stem-loop of african cassava mosaic virus DNA are rapidly corrected by genomic recombination. *Journal of General Virology*, 75(11):3203–3209, 1994.

Rodamilans, B., Shan, H., Pasin, F., and Garca, J. A. Plant viral proteases: Beyond the role of peptide cutters. *Frontiers in Plant Science*, 9:666, 2018.

Roossinck, M. . J. ., Sleat, D., and Palukaitis, P. . Satellite rnas of plant viruses: structures and biological effects. *Microbiol Revs*, 56(3):256–279, 1992.

Roossinck, M. J. Symbiosis versus competition in plant virus evolution. *Nat. Revs. Microbiol.*, 3:917–924, 2005.

Roossinck, M. J. Lifestyles of plant viruses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 365 (1548):1899–1905, 2010.

Roossinck, M. J. *Virus: An Illustrated Guide to 101 Incredible Microbes*. Princeton University Press, New Jersey, 2016.

Rosvall, M. and Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Nat. Acad. Sci.*, 105(4):1118–1123, 2008.

S., H. A. and Baltimore, D. Defective viral particles and viral disease processes. *Nature*, 226(5243):325–327, 1970.

Sachs, J. L. and Bull, J. J. Experimental evolution of conflict mediation between genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (2):390–395, 2005.

Sánchez-Navarro, J., Zwart, M. P., and Elena, S. F. *J. Virol.*, 87:10805, 2013.

Sanjuan, R. Collective infectious units in viruses. *Trends in Microbiology*, 25(5):402 – 412, 2017.

Sano, E., Carlson, S., Wegley, L., and Rohwer, F. Movement of viruses between biomes. *Applied and Environmental Microbiology*, 70(10):5842–5846, 2004.

Saunders, K. and Stanley, J. A nanovirus-like dna component associated with yellow vein disease of ageratum conyzoides: evidence for interfamilial recombination between plant dna viruses. *Virology*, 264(1):142 – 52, 1999.

Scheets, K. Maize chlorotic mottle machlomovirus and wheat streak mosaic rymovirus concentrations increase in the synergistic disease corn lethal necrosis. *Virology*, 242(1): 28 – 38, 1998.

Schmitt, M. J. and Breining, F. The viral killer system in yeast: from molecular biology to application. *FEMS Microbiol Rev*, 26:257–275, 2002.

Scola, B. L., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J.-M., and Raoult, D. A giant virus in amoebae. *Science*, 299(5615):2033–2033, 2003.

Scola, B. L., Desnues, C., Pagnier, I., Robert, C., Barrasi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., and Raoult, D. The virophage as a unique parasite of the giant mimivirus. *Nature*, 455:100–104, 2008.

Selling, B., Allison, R., and P, K. Genomic rna of an insect virus directs synthesis of infectious virions in plants. *Proc Natl Acad Sci USA.*, 87(1):434–438, 1990.

Shackelton, L. A. and Holmes, E. C. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. *Trends Microbiol.*, 12:458–465, 2004.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., Buchmann, J., Wang, W., Xu, J., Holmes, E. C., and Zhang, Y.-Z. Redefining the invertebrate rna virosphere. *Nature*, 540(539), 2016.

Sicard, A., Yvon, M., Timchenko, T., Gronenborn, B., Michalakis, Y., Gutiérrez, S., and Blanc, S. Gene copy number is differentially regulated in a multipartite virus. *Nat. Comm.*, 4:2248, 2013.

Sicard, A., Michalakis, Y., Gutiérrez, S., and Blanc, S. The strange lifestyle of multipartite viruses. *PLoS Path.*, 12:e1005819, 2016.

Sicard, A., Pirolles, E., Gallet, R., Vernerey, M.-S., Yvon, M., Urbino, C., Peterschmitt, M., Gutierrez, S., Michalakis, Y., and Blanc, S. A multicellular way of life for a multipartite virus. *eLife*, 8:e43599, 2019.

Simmonds, P., Adams, M. J., Benkö, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M. Q., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., Sabanadzovic, S., Sullivan, M. B., Suttle, C. A., Tesh, R. B., van der Vlugt, R. A., Varsani, A., and Zerbini, F. M. Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.*, **:1–8, 2017.

Simon-Loriere, E. and Holmes, E. C. Gene duplication is infrequent in the recent evolutionary history of rna viruses. *Mol. Biol. Evol.*, 30:1263–1269, 2013.

Sõmera, M., Sarmiento, C., and Truve, E. Overview on sobemoviruses and a proposal for the creation of the family sobemoviridae. *Viruses*, 7(6):3076–3115, 2015.

Stamos, J. L., Lentzsch, A. M., and Lambowitz, A. M. Structure of a thermostable group II intron reverse transcriptase with template-primer and its functional and evolutionary implications. *Mol. Cell*, 68(5):926 – 939.e4, 2017. ISSN 1097-2765.

Stanley, J., Saunders, K., Pinner, M. S., and Wong, S. M. Novel defective interfering dnas associated with ageratum yellow vein geminivirus infection of ageratum conyzoides. *Virology*, 239(1):87 – 96, 1997.

Stewart, L. R., Ding, B., and Falk, B. W. *Viroids and Phloem-Limited Viruses: Unique Molecular Probes of Phloem Biology*, chapter 13, pages 271–292. John Wiley & Sons, Ltd, 2012.

Stobbe, A. H., Melcher, U., Palmer, M. W., Roossinck, M. J., and Shen, G. Co-divergence and host-switching in the evolution of tobamoviruses. *J. Gen. Virol.*, 93(2):408–418, 2012.

Syller, J. Molecular and biological features of umbraviruses, the unusual plant viruses lacking genetic information for a capsid protein. *Physiological and Molecular Plant Pathology*, 63:35–46, 2003.

Symons, R. H. The intriguing viroids and virusoids: What is their information content and how did they evolve? *Molecular Plant-Microbe Interactions*, 4:111–121, 1991.

Szathmáry, E. Viral sex, levels of selection, and the origin of life. *J. Theor. Biol.*, 159: 99–109, 1992.

Taylor, B. P., Cortez, M. H., and Weitz, J. S. The virus of my virus is my friend: Ecological effects of virophage with alternative modes of coinfection. *J. Theor. Biol.*, 354:124–136, 2014.

Thompson, J. R., Kamath, N., and Perry, K. L. An evolutionary analysis of the secoviridae family of viruses. *PLOS ONE*, 9(9):1–16, 09 2014.

Tilman, D. and Kareiva, P. *The Role of Space in Population Dynamics and Interspecific Interactions*. Princeton University Press, 1997.

Tilsner, J. and Oparka, K. J. Missing links? the connection between replication and movement of plant RNA viruses. *Current Opinion in Virology*, 2(6):705 – 711, 2012.

Timchenko, T., Katul, L., Aronson, M., Vega-Arreguín, J. C., Ramirez, B. C., Vetten, H. J., and Gronenborn, B. Infectivity of nanovirus DNAs: induction of disease by cloned genome components of *faba bean necrotic yellow virus*. *J. Gen. Virol.*, 87:1735–1743, 2006.

Tomasicchio, M., Venter, P. A., Gordon, K. H. J., Hanzlik, T. N., and Dorrington, R. A. Induction of apoptosis in *saccharomyces cerevisiae* results in the spontaneous maturation of tetravirus procapsids *in vivo*. *J. Gen. Virol.*, 88:1576–1582, 2007.

Torrance, L., Lukhovitskaya, N. I., Schepetilnikov, M. V., Cowan, G. H., Ziegler, A., and Savenkov, E. I. Unusual long-distance movement strategies of potato mop-top virus rnas in nicotiana benthamiana. *Molecular Plant-Microbe Interactions*, 22(4):381–390, 2009.

Tromas, N., Zwart, M. P., Lafforgue, G., and Elena, S. F. Within-host spatiotemporal dynamics of plant virus infection at the cellular level. *PLOS Genetics*, 10(2):1–14, 2014.

ul Rehman, M. S. N. and Fauquet, C. M. Evolution of geminiviruses and their satellites. *FEBS Lett.*, 583:1825–1832, 2009.

Valdano, E., Manrubia, S., Gmez, S., and Arenas, A. Endemicity and prevalence of multipartite viruses under heterogeneous between-host transmission. *PLOS Comp. Biol.*, 15 (3):1–21, 03 2019.

van Kammen, A. Purification and properties of the components of cowpea mosaic virus. *Virology*, 31:633–642, 1967.

van Vloten-Doting, L. Advantages of multipartite genomes of single-stranded rna plant viruses in nature, for research, and for genetic engineering. *Plant Molecular Biology Reporter*, 1(2):55–60, 1983.

Varsani, A., Lefeuvre, P., Roumagnac, P., and Martin, D. Notes on recombination and reassortment in multipartite/segmented viruses. *Curr. Op. Virol.*, 33:156 – 166, 2018.

Wang, I.-N., Yeh, W.-B., and Lin, N.-S. Phylogeography and coevolution of bamboo mosaic virus and its associated satellite rna. *Frontiers in Microbiology*, 8:886, 2017a.

Wang, Z., Deng, X., Zou, W., Yan, Z., and Qiu, J. Human bocavirus 1 is a novel helper for adeno-associated virus replication. *J Virol*, 91(18):e00710–17, 2017b.

White, K. Formation and evolution of *tombusvirus* defective interfering rnas. *Seminars in Virology*, 7:409 – 416, 1996.

White, K. and Nagy, P. D. Advances in the molecular biology of iruses: gene expression, genome replication, and recombination. *Progress in Nucleic Acid Research and Molecular Biology*, 78:187 – 226, 2004.

Whitfield, A. E., Falk, B. W., and Rotenberg, D. Insect vector-mediated transmission of plant viruses. *Virology*, 479-480:278 – 289, 2015.

Whitfield, A. E., Huot, O. B., Martin, K. M., Kondo, H., and Dietzgen, R. G. Plant rhabdoviruses–their origins and vector interactions. *Curr. Op. Virol.*, 33:198 – 207, 2018.

Whittaker, G. R. and Helenius, A. Nuclear import and export of viruses and virus genomes. *Virology*, 246(1):1 – 23, 1998.

Wilson, D. S. A theory of group selection. *PNAS*, 72:143–146, 1975.

Wodarz, D. Evolutionary dynamics of giant viruses an their hosts. *Ecol Evol*, 3(7):2103–15, 2013.

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja, V. V., and Koonin, E. V. Origins and evolution of the global RNA virome. *mBio*, 9(6), 2018.

Wu, B., Zwart, M. P., Sánchez-Navarro, J. A., and Elena, S. F. Within-host evolution of segments ratio for the tripartite genome of alfalfa mosaic virus. *Sci Rep*, 7:5004, 2017.

Wylie, S. J. and Jones, R. A. C. Role of recombination in the evolution of host specialization within *bean yellow mosaic virus*. *Phytopathology*, 99:5:512–518, 2009.

Xiong, R., Wu, J., Zhou, Y., and Zhou, X. Identification of a movement protein of the tenuivirus rice stripe virus. *J. Virol.*, 82(24):12304–12311, 2008.

Yau, S., Lauro, F. M., DeMaere, M. Z., Brown, M. V., Thomas, T., Raftery, M. J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J. M., Gibson, J. A., and Cavicchioli, R. Virophage control of antarctic algal host–virus dynamics. *Proc Natl Acad Sci USA*, 108(15):6163–6168, 2011.

Zahid, K., Zhao, J.-H., Smith, N. A., Schumann, U., Fang, Y.-Y., Dennis, E. S., Zhang, R., Guo, H.-S., and Wang, M.-B. Nicotiana small rna sequences support a host genome origin of cucumber mosaic virus satellite rna. *PLOS Genetics*, 11(1):1–13, 01 2015.

Zhang, X.-F. and Qu, F. Multi-component plant viruses. *eLS*, pages 1–7, 2015.

Zhang, Y.-J., Wu, Z.-X., Holme, P., and Yang, K.-C. Advantage of being multicomponent and spatial: Multipartite viruses colonize structured populations with lower thresholds. *Phys. Rev. Lett.*, 123:138101, 2019.

Zhong, B.-X., Shen, Y.-W., Omura, T., and Shen, D.-L. RNA-binding Domain of the Key Structural Protein P7 for the Rice dwarf virus Particle Assembly. *ABBS*, 37(1):55–60, 2005.

Zwart, M. P. and Elena, S. F. Matters of size: genetic bottlenecks n virus infection and their potential impact on evolution. *Annu. Rev. Virol.*, 2:161–179, 2015.

Zwart, M. P., Willemsen, A., Dars, J.-A., and Elena, S. F. Experimental Evolution of Pseudogenization and Gene Loss in a Plant RNA Virus. *Mol. Biol. Evol.*, 31(1):121–134, 2014.

APPENDIX A

SCRIPTS

A.1 C scripts

A.1.1 Model of genome segmentation

An implementation of stochastic simulations in C language are used to solve the model of genome segmentation presented in Chapter 2. The script is given below:

Listing A.1: C script of genome segmentation model

```
/*
        segmentspace_8.c - 01/06/16 A Sanz
                Is equivalent to segmentspace_1.c but changing some things
                the data.
                In order to decrease the computation time we only need prev
                Lets transform M the matrix that stores the viral types, in
                M(i,x,t)->M(i,x) "independent" from time.
                Two matrices. M and M_

        -STEPS:
                1- Infect. MOI
                2- Segment. MU
                3- Complement. min condition
```

```
        -COMPILATION: gcc -std=c99 -o runable segmentspace.c -lm
        -FILE I/O: Three files are produced (one for each viral type). Change the paths
                    purposes in the section "substitution variables".

        -OTHER NOTES
                BUG RESOLVED. [Fi/sum(Fi)=nan] if you have empty cells. In this case yo
                CAUTION. THe matrix u should save is M after infection. Otherwise you w
                        conserve the number of virus per cell.



*/

#include <stdio.h>      //standard library
#include <stdlib.h>     // RAND_MAX to have a distribution from 0 to 1
#include <time.h>       // NULL to change seed
#include <math.h>       // typical functions

// DIRECTLY SUBSTITUTION VARIABLES

#define TYPES 3
#define TIME 1000
#define CELLS 100
#define MU 0.1
#define MOI 2

#define PATH1 "/home/newton/Documentos/Introd_C_programming/some-figures/space_wt_im4.d
#define PATH2 "/home/newton/Documentos/Introd_C_programming/some-figures/space_a_im4.da
#define PATH3 "/home/newton/Documentos/Introd_C_programming/some-figures/space_b_im4.da

// HEAD: FUNCTION DECLARATION


void segment( int [][CELLS] ); //input a matrix by reference.
void infect(int [][CELLS], int [][CELLS]);
void zeros( int [][CELLS], int);
void complement( int [][CELLS],int [][CELLS]);
int sumc(int [][CELLS], int);
int sum(int [][CELLS], int);
int min(int, int);
double rand_uni();



// START THE COMPUTATION HERE

int main(){
        //Seed is started with clock time.
        srand(time(NULL));
```

```
//Variables
// Files to save data. One for each viral specie.
FILE *wt,*A,*B;
wt=fopen(PATH1,"w");
A=fopen(PATH2,"w");
B=fopen(PATH3,"w");
// Variables
int WT_0=MOI;
int M[TYPES][CELLS];
int M_[TYPES][CELLS];

// Two ways to initialize the matrix with wt virus
for (int x=0;x<CELLS;x++){
        zeros(M,x);
        zeros(M_,x);
        //1. Random cells are occupied at the beginning
        /*if (rand_uni()<0.2){
                M_[0][x]=WT_0;
        }*/
        //2. All cells occupied at the beginning
        M_[0][x]=WT_0*100;
}

//Start the spatio-temporal dynamics
for (int t=1;t<TIME; t++ ){

        infect(M,M_);



        //Uncomment this to save the spatio-temporal----//
        // species after infection                      //
        for (int x=0;x<CELLS;x++){                      //
                fprintf(wt,"%d ",M[0][x]);              //
                fprintf(A,"%d ",M[1][x]);               //
                fprintf(B,"%d ",M[2][x]);               //
        }                                               //
        fprintf(wt,"\n");                               //
        fprintf(A,"\n");                                //
        fprintf(B,"\n");                                //
        //-----------------------------------------------//


        /*//Uncomment this to save only temporal        --------//
        //      variation after infection               //
        //      (Sums up the virus)                     //
                                                        //
        fprintf(wt,"%d\n",sumc(M,0));                   //
        fprintf(A,"%d\n",sumc(M,1));                    //
```

```
                fprintf(B,"%d\n",sumc(M,2));                    //
                                                                //
               //-------------------------------------------------//
                */
                segment(M);
                complement(M,M_);

        }

        //close files
        fclose(wt);
        fclose(A);
        fclose(B);

        return 0;
}

//FUNCTIONS

/*
        "Complement" takes the matrix M where the virus are storaged and applies the mi
        for each cell. Returns matrix M_ with the result and also initializes with zero
*/

void complement( int M[][CELLS], int M_[][CELLS]){
        for ( int j=0;j<CELLS;j++ ){
                //Complementation
                M_[0][j]=M[0][j];
                M_[1][j]=min(M[1][j],M[0][j]+M[2][j]);
                M_[2][j]=min(M[2][j],M[0][j]+M[1][j]);
                zeros(M,j);
        }
}

/*
        "infect" takes the virus in matrix M_ (past) and infects cells of matrix M (pre
        2 neigbouring parental cells. Uses a poisson distribution of the virus particle
*/

void infect(int M[][CELLS], int M_[][CELLS] ){
        double p1,p2,p3,f,rr;

        for ( int x=0; x<(CELLS-1); x++ ){
                zeros(M,x);
                for ( int i=0; i<MOI; i++ ){
                        rr=rand_uni();
                        f=0.0;
                        for ( int k=0; k<TYPES; k++ ){
                                p1=M_[k][x]+M_[k][x+1];
```

```
                                        p2=sum(M_,x)+sum(M_,x+1);
                                        p3=p1/p2;

                                        if (isnan(p3)){
                                                p3=0;

                                        }
                                        if ( (f<=rr) && rr<=(p3+f) ){
                                                M[k][x]+=1;// Add one virus to cell
                                        }
                                        f=f+p3;
                                }
                        }
                }

                // Periodic contour conditions

                zeros(M,CELLS-1);
                for ( int j= 0; j<MOI; j++ ){
                        rr=rand_uni();
                        f=0.0;
                        for ( int q=0; q<TYPES;q++ ){
                                p1=M_[q][CELLS-1]+M_[q][0];
                                p2=sum(M_,CELLS-1)+sum(M_,0);
                                p3=p1/p2;
                                if (isnan(p3)){
                                        p3=0;

                                }
                                if ( (f<=rr) && rr<=(p1/p2+f) ){
                                        M[q][CELLS-1]+=1;
                                }
                                f=f+p3;
                        }
                }
        }

        /*      "segment" takes the matrix M where virus are storaged and
                randomly fragments the wt type with a binomial probabilty MU (chang
                in the section "direct substitution variables")
        */


        void segment(int M[][CELLS]){
                int k;
                double r;
                for ( int x=0;x<CELLS;x++ ){
                        //take the number of wt virus from time t --> k for all cel
```

```
                k=M[0][x];
                while ( k>0 ){
                        k--;
                        r=rand_uni();
                        if ( r<=MU ){
                                M[0][x]-=1;      //wt
                                M[1][x]+=1;      //a
                                M[2][x]+=1;      //b
                        }
                }
        }
}

/*      "rand_uni" Returns a uniform random number from 0 to 1
*/
double rand_uni(){return rand()/(double)RAND_MAX;}


/*      "zeros" fills with zeros the column "column" of any matrix "M"
        Need to specify the total number of columns "tot_columns"
*/
/*Initialize the matrix M with zeros for all the species in the input columns*/
void zeros(int M[][CELLS], int cell){

        for ( int l=0;l<TYPES;l++ ){M[l][cell]=0;}
}



/*      "sum" all the virus types wt+a+b stored in "M" at a given time "time"
        This is equivalent to the MATLAB function sum(M(:,t)).
        Returns a variable that is the sum
*/
int sum(int M[][CELLS],int cell){
        int s=0;
        for ( int j=0;j<TYPES;j++ ){
                s+=M[j][cell];
        }

        return s;
}

/*
        sumc is the same as sum but this time sums the virus in all the cells
        instead of the virus of all types in each cell.
*/
int sumc(int M[][CELLS], int cell){
        int s=0;
        for ( int j=0; j<CELLS; j++){s+=M[cell][j];}
```

```
        return s;
}

/*      min(a,b) gives the minimun integrer between a and b.
*/

int min( int A, int B){
        if ( A<B ){return A;}
        else{return B;}
}
```

A.1.2 Model of viral competition assisted by a satellite

The equations of model of two virus competing for infection with a satellite presented in Chapter 3 are numerically solved in C using the algorithm Runge-Kutta. The script is given below:

Listing A.2: C script of competition assisted by a satellite

```
/*
        Adriana Sanz 22/05/2017
   satcompet.c is a simple model of viral competition.

        Implements Runge Kutta integration step.
        Fist order system of equations:
                h' = g - (d +p_x*x+p_y*y + p_sy*s)*h
                x' = (p_x*h - (d + d_x))*x
                y' = (p_y*h - (d + d_y) - p_s*s)*y
                s' = (p_s*y -(d + d_s) + p_sy*h)*s
        Variables
                h heathy hosts
                x infected hosts by virusx
                y infected hosts by virusy
        Parameters
                g linear growth rate
                p_i prob of get infected by i E {x,y}
                d_i prob of increased death being infected by i E {x,y}

compilation routine: gcc -std=c99 sat.c -o sat

INCLUDES change of parameters

*/
//////////////////////////////////////
// HEADER
//////////////////////////////////////

# include <stdio.h>
# include <stdlib.h>
# include <time.h>
```

```
//////////////////////////////////////
// DEFINITIONS DECLARATIONS
//////////////////////////////////////

#define TIME 250000
#define PATH1 "psy_dx_x_y.dat"
#define PATH2 "psy_dx_y_y.dat"
#define PATH3 "psy_dx_s_y.dat"
#define PATH4 "psy_dx_h_y.dat"
#define U 100
#define V 100

double dh(double *, double ,double ,double, double);
double dx(double *, double ,double ,double);
double dy(double *, double ,double ,double, double);
double ds(double *, double ,double ,double, double);
void rk(double *, double *, double ,double ,double, double );
void integration_rk(double [][4], double *, double *, double *, double);
double sum(double *);
void linspace(double *,double, double, int);

//////////////////////////////////////
// START
//////////////////////////////////////
int main(){

//parameters
        double initial[]={1,0.2,0.2,1}; //vector for initial conditions
        double final[4];                //vector for final conditions
        double k[4][4];                 //vector to calculate runge kutta
        double p[U];
        double q[V];
        linspace(p,0.0,1.,V);
        linspace(q,0.0,1.,U);

        double param[10];// y wins      x wins
        param[0]= 0.2;   // g=1         1
        param[1]= 0.05;  // d=0.05      0.05
        param[2]= 0.2375;// p_x=0.2     0.3
        param[3]= 0.1;   // p_y=0.3     0.2
        param[4]= 0.25;  // d_x=0.1     0.01
        param[5]= 0.1816;// d_y=0.01    0.01
        param[6]= 1;  // p_s=0.8        0.25
        param[7]= 1.25;  // p_sy=0.25   0.25
        param[8]= 0.54512;// d_s=0.3    0.01
        param[9]= 0.0227;// p_Y=0.3     0.25
```

```
        double tstep=0.01; // time step

//////////////////////////////////////
// File IO
//////////////////////////////////////
        FILE *fp1, *fp2, *fp3, *fp4;            // create a pointer "fp" to
        fp1=fopen(PATH1,"w");    // indicate where to save/name
the file and the option "w"
        fp2=fopen(PATH2,"w");
        fp3=fopen(PATH3,"w");                   // indicate write
        fp4=fopen(PATH4,"w");
//SET TIME
        time_t start, stop;
        time(&start);
        for (int i=0; i<V; i++){        // initial loops for change of para
                param[4]=q[i];
                for (int u=0;u<U;u++){
                        initial[0]=1;
                        initial[1]=0.000001;
                        initial[2]=1;
                        initial[3]=1;
                        param[7]=p[u];

                        for (int t=0;t<TIME;t++){       //temporal loop

                                //integrate runge kutta for each time
                                integration_rk(k,param,initial,final,tstep)


                                //change the intial value
                                initial[0]=final[0];
                                initial[1]=final[1];
                                initial[2]=final[2];
                                initial[3]=final[3];
                                if (t==TIME-1){
                                        fprintf(fp1,"%.10f\t",final[1]);
                                        fprintf(fp2,"%.10f\t",final[2]);
                                        fprintf(fp3,"%.10f\t",final[3]);
                                        fprintf(fp4,"%.10f\t",final[0]);
                                }
                        }
                        //printf("iteration t=%d,s_0=%f\n",u,ds[u]);
                }
                printf("iteration t=%d,dx=%f\n",i,p[i]);
                fprintf(fp1,"\n");
                fprintf(fp2,"\n");
                fprintf(fp3,"\n");
                fprintf(fp4,"\n");
        }
```

```
        fclose(fp1);
        fclose(fp2);
        fclose(fp3);
        fclose(fp4);
        time(&stop);
        printf("\nElapsed time: %.0f seconds for %d iterations \n", difftime(stop, star

        return 0;
}
//END

//////////////////////////////////////
//    FUNCTIONS
//////////////////////////////////////
void integration_rk(double k[][4], double *variables, double *initial,double *final,dou

        double x_0,y_0,h_0,s_0;
        h_0=initial[0];
        x_0=initial[1];
        y_0=initial[2];
        s_0=initial[3];

        rk(k[0],variables,h_0,x_0,y_0,s_0);
        //printf("k1=%f k2=%f k3=%f k4=%f\n",k[0][0],k[0][1],k[0][2],k[0][3]);
        rk(k[1],variables, h_0 + (tstep/2)*k[0][0], x_0 + (tstep/2)*k[0][1], y_0 + (tst
        //printf("k1=%f k2=%f k3=%f k4=%f\n",k[1][0],k[1][1],k[1][2],k[1][3]);
        rk(k[2],variables, h_0 + (tstep/2)*k[1][0], x_0 + (tstep/2)*k[1][1], y_0 + (tst
        //printf("k1=%f k2=%f k3=%f k4=%f\n",k[2][0],k[2][1],k[2][2],k[2][3]);
        rk(k[3],variables, h_0 + tstep*k[2][0], x_0 + tstep*k[2][1], y_0 + tstep*k[2][2
        //printf("k1=%f k2=%f k3=%f k4=%f\n",k[3][0],k[3][1],k[3][2],k[3][3]);
        final[0]= h_0 + tstep*(k[0][0]+2*k[1][0]+2*k[2][0]+k[3][0])/6;
        final[1]= x_0 + tstep*(k[0][1]+2*k[1][1]+2*k[2][1]+k[3][1])/6;
        final[2]= y_0 + tstep*(k[0][2]+2*k[1][2]+2*k[2][2]+k[3][2])/6;
        final[3]= s_0 + tstep*(k[0][3]+2*k[1][3]+2*k[2][3]+k[3][3])/6;
        //printf("h=%f x=%f y=%f s=%f\n",final[0],final[1],final[2],final[3]);
}

void rk(double v[4], double p[8], double h,double x,double y,double s){

        v[0]=dh(p,h,x,y,s);
        v[1]=dx(p,h,x,s);
        v[2]=dy(p,h,x,y,s);
        v[3]=ds(p,h,x,y,s);
        //printf("h=%f x=%f y=%f s=%f\n--\n",v[0],v[1],v[2],v[3]);
}

double dh(double *v,double h,double x,double y,double s){
        double g=v[0];
        double d=v[1];
```

```
        double p_x=v[2];
        double p_y=v[3];
        double p_sy=v[7];
        double p_Y=v[9];
        return g - (d + p_x*x + p_y*y + (p_sy+p_Y)*s)*h;
}
double dx(double *v,double h,double x,double y){
        double d=v[1];
        double p_x=v[2];
        double d_x=v[4];
        return (p_x*h-(d+d_x))*x;
}
double dy(double *v,double h,double x,double y,double s){
        double d=v[1];
        double p_y=v[3];
        double d_y=v[5];
        double p_s=v[6];
        double p_Y=v[9];
        return p_y*y*h - (d+d_y)*y - p_s*s*y + p_Y*s*h;
}
double ds(double *v,double h,double x,double y, double s){
        double d=v[1];
        double p_s=v[6];
        double p_sy=v[7];
        double d_s=v[8];
        return p_s*y*s -(d + d_s)*s + p_sy*h*s;
}


void linspace(double *v,double MIn, double MAx, int LENGTH){
        double s=(MAx-MIn)/(LENGTH-1);
        v[0]=MIn;
        for ( int i =1; i< LENGTH; i++ ){v[i]=s+v[i-1];}
}
double sum(double *v){
        double s=0;
        for ( int j=0;j<4;j++ ){s+=v[j];}
        return s;
}
```

A.2 Matlab functions

In combination with the bioinformatics toolbox of MATLAB several functions to retrieve information from a phylogenetic tree.

A.2.1 Calculate evolutionary distances in a phylogenetic tree

We start finding in the phylogenetic tree the leaves of interest where a particular trait is located. For example, the leaves were multipartite and segmented species are located. The function generate a zeros-and-ones vector with the leaves of interest.

Listing A.3: Find the leaves were leaves of interest are located

```
function foundLeaves=findLeaves(numLeaves,ID,list)
% A=numLeaves;
B=length(list);
foundLeaves=zeros(numLeaves,1);
for i=1:numLeaves
    a=char(ID(i));
    c=str2num(a);
%     display(c);
    for j=1:B
%         b=char(list(j));
        b=list(j);
%         display(b);
        if c==b
            foundLeaves(i)=1;
        end
    end
end
```

Normally, leaves that share a trait are grouped forming cluster. Therefore, we next find the internal branches or nodes where groups of species of interest emerged in the phylogenetic tree. Unique or isolated species or leaves are excluded from the analysis. The function generate a zeros-and-ones vector where the branches where the traits emerged.

Listing A.4: Find the branches of origin of the trait

```
function [mutantAncestor_in,leaf_out]=findBranchMutated(leaf_in,NodeOrder,numLeaves,Mat
mutantAncestor_out=find(Matrix(:,leaf_in));

mutSubtree=GetSubTreeNodes(mutantAncestor_out,Matrix);

mutantLeaves=mutSubtree(find(NodeOrder(mutSubtree)==0));

curr_state=sum(listOfmutants(mutantLeaves))/length(mutantLeaves);

if curr_state~=1
    leaf_out=leaf_in +1;
    mutantAncestor_in=leaf_in;
else
    while curr_state==1.
        mutantAncestor_in=mutantAncestor_out;
        mutantAncestor_out=find(Matrix(:,mutantAncestor_in));% find ancestor
        mutSubtree=GetSubTreeNodes(mutantAncestor_out,Matrix);% get the subtree of ance
        mutantLeaves=mutSubtree(find(NodeOrder(mutSubtree)==0)); %check the leaves
        curr_state=sum(listOfmutants(mutantLeaves))/length(mutantLeaves); % will be 1 i
```

```
    end
    mutSubtree=GetSubTreeNodes(mutantAncestor_in,Matrix);% get the subtree
    mutantLeaves=mutSubtree(find(NodeOrder(mutSubtree)==0)); %check the lea
    leaf_out=max(mutantLeaves)+1;
end
end
```

We can calculate the distance from those branches to the MRCA that is the root of the phylogenetic tree.

Listing A.5: Get the distance to the root or MRCA

```
function distroot=GetDistance2root(node,Distances,Matrix)
state = find(Matrix(:,node));
distroot=Distances(node);
while ~isempty(state)
    distroot=distroot + Distances(state);
    state=find(Matrix(:,state));
end
end
```

Also we can calculate the weighted distance to the leaves from the branches. Not-ultrametric trees have a different length for each leaves.

Listing A.6: Get the weighted distance to the leaves

```
function H0_out = CalculateH0(Distances,Matrix,NodeOrder,IX_node)

if NodeOrder(IX_node)==0
    H0_out = Distances(IX_node);
else
    IX_subtree = GetSubTreeNodes(IX_node,Matrix);
    [~,IX_order] = sort(NodeOrder(IX_subtree));% uses only the indexes
    IX_subtree = IX_subtree(IX_order);
    for i_node = 1:length(IX_subtree)
        IX = IX_subtree(i_node);

        B0 = Distances(IX);
        IX_children = find(Matrix(IX,:));
        if isempty(IX_children)
            H0(IX) = B0; % this is a leaf
        else
            IX1 = IX_children(1);
            IX2 = IX_children(2);

            IX1_subtree1 = GetSubTreeNodes(IX1,Matrix);
            IX2_subtree2 = GetSubTreeNodes(IX2,Matrix);

            H1 = H0(IX1);
            W1 = sum(Distances(IX1_subtree1));

            H2 = H0(IX2);
```

```
            W2 = sum(Distances(IX2_subtree2));

            H0(IX) = B0 + (W1*H1 + W2*H2)/(W1+W2);
        end
    end
    H0_out = H0(IX_node);
end
```

Finally, we can calculate the branching distance. The branching distance is the evolutionary distance of two sequences that have diverged.

Listing A.7: Get subtree nodes

```
function IX_out = GetSubTreeNodes(IX_branch,Matrix)
IX_out = [];
IX_currentStage = IX_branch;
% k_temp = 1;
while ~isempty(IX_currentStage)
    IX_NextStage = [];
    for i_node = 1:length(IX_currentStage)
        IX_temp = find(Matrix(IX_currentStage(i_node),:));
        IX_out = [IX_out IX_temp];
        IX_NextStage = [IX_NextStage IX_temp];
    end
    IX_currentStage = IX_NextStage;
end

IX_out = [IX_out IX_branch];
```

Listing A.8: Get the branching distance

```
function [ancestors]=BranchDistances(NodeOrder,Matrix,listOfmutants)
mutant_nodes=find(listOfmutants);
numMutants=sum(listOfmutants);
Ancestor_nodes=[];
index=1;
index_mut=mutant_nodes(index);

while index<=numMutants
    ancestor_mut=find(Matrix(:,index_mut));
    subtree_mut=GetSubTreeNodes(ancestor_mut,Matrix);
    mutantLeaves=sort(subtree_mut(find(NodeOrder(subtree_mut)==0)));

    if (sum(listOfmutants(mutantLeaves))/length(mutantLeaves))==1
        index=max(find(listOfmutants(mutantLeaves)==1))+1;
        index_mut=ancestor_mut;
        Ancestor_nodes=[Ancestor_nodes subtree_mut(find(NodeOrder(subtree_mut)~=0))];
    else
        index=max(find(listOfmutants(mutantLeaves)==1))+1;
%         index_mut=mutant_nodes(index);
            if index>length(mutant_nodes)
```

```
                index=length(mutant_nodes);
                index_mut=length(listOfmutants);
            else
                index_mut=mutant_nodes(index);
            end

    end
    if rem(index_mut,10)==0
        fprintf('ancestor=%d,leave=%d\n',ancestor_mut,max(find(listOfmutant
    end
end
ancestors=unique(Ancestor_nodes);
```